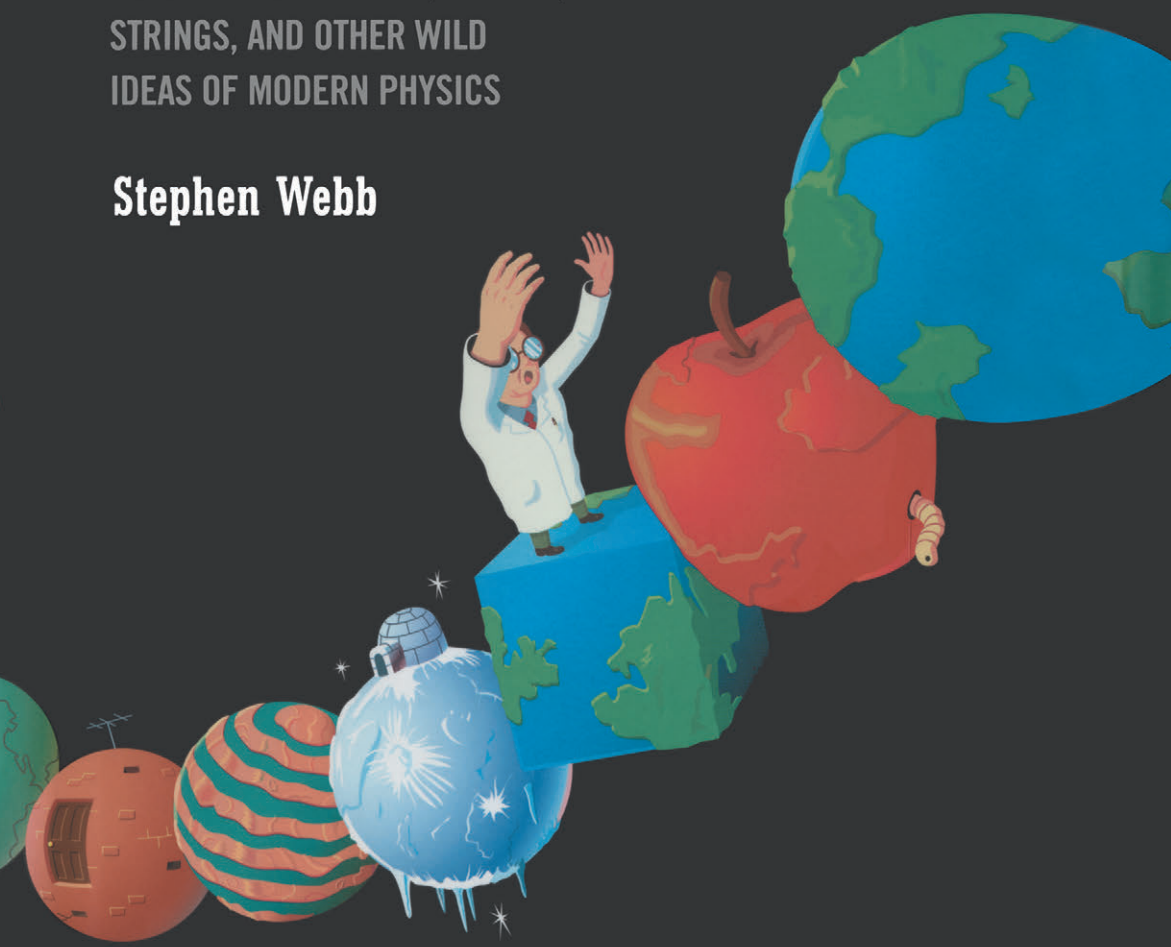
OUT OF THIS WORLD

COLLIDING UNIVERSES, BRANES, STRINGS, AND OTHER WILD IDEAS OF MODERN PHYSICS

Stephen Webb



OUT OF THIS WORLD

OUT OF THIS WORLD

COLLIDING UNIVERSES, BRANES, STRINGS,
AND OTHER WILD IDEAS OF MODERN PHYSICS

Stephen Webb

Springer-Verlag Berlin Heidelberg GmbH

© 2004 Springer-Verlag Berlin Heidelberg
Originally published by Springer-Verlag Berlin Heidelberg New York in 2004
Softcover reprint of the hardcover 1st edition 2004

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher.

www.copernicusbooks.com

Library of Congress Cataloging-in-Publication Data
Webb, Stephen
Out of this world : colliding universes, branes, strings, and other wild ideas of modern physics / Stephen Webb.
p. cm.
Includes index.
ISBN 978-1-4419-1852-9 ISBN 978-1-4757-6120-7 (eBook)
DOI 10.1007/978-1-4757-6120-7
1. Physics—Popular works. I. Title.
QC24.5.W43 2004
539—dc22

2004041822

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

ISBN 978-1-4419-1852-9 SPIN 10926451

To Jessica

CONTENTS

INTRODUCTION		1
	Gedanken Physics	2
	The Allegory of the Cave	4
	The Search for Symmetry	7
CHAPTER 1:	SYMMETRY	9
	Symmetry in Everyday Life	11
	Snowflakes	15
	Symmetry in Physics	17
	Groups: The Mathematics of Symmetry	25
CHAPTER 2:	TWIN PILLARS OF PHYSICS	29
	The Global Positioning System	30
	Physics on a Large Scale: The Theory of General Relativity	34
	Physics on a Small Scale: The Weird World of the Quantum	39
	When GR Meets QM: Big Trouble at Little Distances	46
CHAPTER 3:	BUILDING BLOCKS OF THE UNIVERSE	49
	Producing the Fundamental Particles	51
	Classifying the Fundamental Particles	59
	A First Look at the Fundamental Particles	64
CHAPTER 4:	THE STANDARD MODEL	79
	Our Best Theory	80
	Local Gauge Symmetry	83
	The Electromagnetic Interaction	85
	The Strong Interaction	89
	The Weak Interaction	95
	Electroweak Unification and the Higgs	99
	The Successes of the Standard Model	107
	The Standard Model is not a Final Theory	109
CHAPTER 5:	GUTS AND SUSY	113
	Grand Unification Theories	114
	Supersymmetry	127
	Local Supersymmetry	134
CHAPTER 6:	HIGHER DIMENSIONS	141
	Kaluza-Klein Theory	142
	Yet More Dimensions	149

CHAPTER 7: SUPERSTRINGS 153
The Early History of Superstrings 156
The First Superstring Revolution 165
The Fruits of the Revolution 169

CHAPTER 8: THE STORY OF M 181
Electromagnetic Duality 182
Duality in Strings 195
M-Theory 201
Conclusion 205

CHAPTER 9: OF BRANES AND BLACK HOLES 207
Branes 208
D-Branes 212
Black Holes 215

CHAPTER 10: OUR WORLD AS A HOLOGRAM 225
Entropy 226
The Holographic Principle 232
The AdS/CFT Correspondence 235
Tying Strings Together: AdS/CFT and Large N 241

CHAPTER 11: WHEN WORLDS COLLIDE 245
Brane New World 247
Probing the Higher Dimensions 257
Why Two Branes are Better than One:
The Randall-Sundrum Scenario 269
Is Spacetime Really Symmetric? 274
The Ekpyrotic Scenario: When Worlds Collide 278
The Anthropic Principle 283
Coming Attractions 284

GLOSSARY 287
BIBLIOGRAPHY 301
INDEX 303

PREFACE

This book is an attempt to share some of the fascinating and exciting ideas of modern theoretical physics with a non-mathematical audience. I also hope to give some appreciation of the exceptionally creative people who have generated these ideas. I had no intention of writing a comprehensive history of these ideas, however; I apologize to those physicists whose important contributions I may have omitted.

Several people have contributed directly to the writing of this book, and I would like to take this chance to thank them. Clive Horwood of Praxis Publishing and Paul Farrell of Copernicus Books have both been supportive of the project. Anna Painter has been an exceptionally thorough editor and has provided invaluable advice. Dr John Mason and Lyman Lyons both offered constructive criticism of an early draft. The errors that remain — and errors and misinterpretations are inevitable in any attempt to explain the subtle ideas of theoretical physics without the aid of mathematics — are my sole responsibility.

I would like to thank Ron, Ronnie, Peter, Jackie, Emily, and Abigail for their support. And, most importantly, I would like to thank Heike and Jessica for their patience.

Stephen Webb
Milton Keynes, April 2004

FIGURE CREDITS

I would like to acknowledge the following sources and copyright holders for granting permission to use their images.

Figure 8 is from the Wilson A. Bentley collection. Figure 14 is courtesy of NASA.

Several particle physics and cosmic-ray laboratories kindly gave permission to use images. DESY, Hamburg, gave permission for Figures 18, 25, and 37; HiRes2, Utah, for Figure 19; CERN, Geneva, for Figures 20, 21, 22, 24, 30, 31, 32, 34, 35, 36, 45, 46, 54, 143, and 131. Fermilab Visual Media Services gave permission to reproduce Figure 23, and the Kamioka Observatory, Institute for Cosmic Ray Research (ICRR), the University of Tokyo, gave permission to reproduce Figure 62. Figure 110 is courtesy of the European Space Agency; Figure 129 is courtesy of the Anglo–Australian Observatory; and Figure 130 is courtesy of the Pierre Auger Observatory.

Several universities and conference organizers also gave permission to use images. Figure 41 is courtesy of the Archives, California Institute of Technology. The Office of Communications, Princeton University, gave permission for Figures 72 and 142; University of California, Santa Barbara for Figure 82; Heriot-Watt University for Figure 91; Boston University for Figure 56; Hong Kong University of Science and Technology for Figure 94; and the Institute of Theoretical Physics, Utrecht, for Figure 27. Figure 139 is courtesy of Indiana University; Figures 75 and 134 were taken at the *Second Meeting on CPT and Lorentz Symmetry*, Indiana University, Bloomington, 2001, organized by Professor Alan Kostelecký; all rights reserved. Figure 73 was taken at the Cremmer–Gervais Fête held in Paris, 2002; all rights reserved. Figure 67 was taken at the *Supergravity at 25* conference held at the State University of New York at Stony Brook, 2001; all rights reserved.

Four companies were kind enough to grant permission to reproduce images: Figure 15 is courtesy of Garmin Ltd, Figure 113 is courtesy of Lucent Technologies, and Figure 115 is courtesy of Universal Hologram. Figure 119, M. C. Escher's *Circle Limit IV*, is © 2004 The M. C. Escher Company – Baarn – Holland; all rights reserved.

Finally, I would like to thank several individuals for granting permission to use their photographs and for responding so readily to email requests. I would like to thank Professor Sir Michael Berry for permission to reproduce Figure 92; Almut Mecke for Figures 98 and 107; Professor John Vander Velde for Figure 61; Professor Oleg Lavrentovich

for Figure 104; Peter Ginter for Figure 33; Mary Levin for Figure 133; Peter Tuffy, Edinburgh University for Figure 52; Professor Burt Ovrut for Figure 140; Professor Neil Turok for Figure 141; and Philippe Plailly/Eurelios for Figure 132. Figure 53 was drawn by Georges Boixader, based on an original idea by Professor David Miller. (Visit http://www.hep.ucl.ac.uk/~djm/higgsa.html for the original version.) Gary Friedman gave permission to reproduce Figure 99. (See The Friedman Archives at http://www.friedmanarchives.com.) Permission to reproduce Figure 83 was kindly given by Jean-François Colonna. (See http://www.lactamme.polytechnique.fr.) Dr Lubos Motl kindly gave permission to reproduce Figures 57, 59, 79, 86, 87, 88, 97, 102, 112, 117, 118, 121, 123, 125, 127, 136, and 137. (Visit http://schwinger.harvard.edu/~motl/PeoplePixWeb/ to see more photographs of string theorists taken by Lubos Motl.)

INTRODUCTION

The realities of the world affected me as visions, and as visions only,
while the wild ideas of the land of dreams became, in turn,
not the material of my every-day existence,
but in very deed that existence utterly and solely in itself.
Edgar Allan Poe, *Berenice*

This book is about some really wild ideas and why a group of sober, professional, extremely intelligent people think there is a chance of these ideas being true.

The ideas have their roots in particle physics — the search for simplicity, order, and the ultimate building blocks from which the Universe is made. The search has led physicists to ever-higher energies and ever-smaller distances. It has led them to consider conditions that occurred in the first few moments after the Big Bang. So the quest to discover what our Universe is made of may also tell us how our Universe started and what its fate might be. There is thus a pleasing mix of simplicity and profundity in these ideas.

The ideas are more imaginative and more outrageous than anything to be found in the pages of modern novels. In the words of the wonderful writer Bill Bryson, they sound "worryingly like the sort of thoughts that would make you edge away if conveyed to you by a stranger on a park bench." The ideas are also more beautiful than what most people would associate with a subject like physics. A prime driver of developments in particle physics is the uncovering of symmetry, or beauty, in the equations. Furthermore, the way in which theoretical physics is now performed — with scholars from around the world contributing to the subject via the internet — has been likened to an intricate and beautiful jazz improvisation. These ideas — imaginative, profound, and beautiful as they are — should be viewed on a par with the contributions that art, music, and literature

make to our society. But there is a crucial difference. *We will soon be able to test some of these amazing contributions in experiments.* In a few years we may know whether these ideas are correct as well as imaginative. It is certain that many of them will turn out to be wrong — that is the usual fate of speculative ideas in science. But if we find evidence for just some of them, our view of the world we inhabit will change forever.

In the final chapter of the book we look at some forthcoming experiments that will search for hidden dimensions, branes, and colliding universes. In the absence of such tests to date, however, theoretical physicists have been forced to use a different type of experiment — the *gedanken* experiment — which shall feature heavily in our discussions. So let's begin by defining "gedanken."

GEDANKEN PHYSICS

Thinking physics is gedanken physics.
Lewis C. Epstein

Gedanken, or thought, experiments are tests of hypotheses that take place in the imagination rather than the laboratory. They are invaluable in cases where it would be impractical or unethical to perform a "real" experiment. Gedanken experiments are discussed in various fields of science. In cognitive science, for example, philosophers debate the zombie question: What would the world be like if normal human beings were replaced by zombies — beings that look and behave like the people we know and love, but who lack conscious experience and for whom all is "dark" inside? An examination of this question may give us clues about the nature of consciousness. In quantum mechanics, physicists study the plight of Schrödinger's cat — an unfortunate creature which, through the quantum phenomenon of the superposition of states, seems as if it must be at once both dead and alive. Understanding the fate of the cat may help us understand how the haze of quantum possibilities switches to a definite state that we can observe. Of all scientists it is perhaps theoretical physicists who are fondest of conducting gedanken experiments, and without doubt it was Albert Einstein who was the greatest exponent of the gedanken experiment.

As a young man, Einstein asked himself what a light wave would look like if he could travel alongside it at the speed of light. He could not perform the experiment directly, of course; instead, he had to rely on principles of symmetry, logic, and mathematical consistency. His arguments also had to be consistent with the rest of physics. After ten years of thought he arrived at an answer. His chain of reasoning had led him to the theory

of special relativity, one of the cornerstones of modern physics. An even more profound result came about when Einstein analyzed the following gedanken experiment. Is there any way for an observer who is enclosed within a small, empty, opaque box to tell for certain whether the force she feels is due to gravity, or to an acceleration? His analysis of this question led him directly to the theory of general relativity, perhaps the most beautiful of all theories in modern physics.

By developing and refining ever-more intricate gedanken experiments, theoretical physicists have created fundamental theories that are astonishingly wide-ranging and successful. In fact, these theories are so wide-ranging that some writers have argued that physicists may be on the verge of writing down a "theory of everything." I will not use the phrase "theory of everything" in this book, since it seems to me to be an incorrect term: whatever turns out to be the ultimate, unified theory of particle physics, it will probably not be useful in explaining, say, the phenomenon of human consciousness. But I *will* often use the terms "fundamental theory" and "fundamental physics." These terms have a particular meaning, and it is worth examining the meaning in a little detail.

Steven Weinberg, a Nobel laureate, wrote a beautiful essay entitled "On a piece of chalk" (the title was taken from a famous lecture of the same name given by Thomas Huxley in the 1860s). In his essay, Weinberg points out that, whenever we ask "why?" questions in science, we seek an answer in the simplest and most economic terms. We can keep on asking "why?" questions, and we end up with what might be called an arrow of explanation. For example: *Why is chalk white?* Well, we can give an answer in terms of the light absorbtion properties of chalk. *Why does chalk have its particular light absorbtion properties?* Well, the answer has to do with the states of the molecules from which chalk is made. If we keep on asking "why?" questions in this way, the arrow of explanation points to the ideas of particle physics (and then to the frontier of unanswered questions in particle physics). The point is that, *wherever we start from*, we end up following an arrow of explanation that leads to the ideas of particle physics. (Richard Feynman, another Nobel laureate, once pointed out something similar in a fascinating television interview. He asked the viewers to imagine something along the lines of the following dialogue. "Aunt Nellie is in the hospital." *Why?* "Well, she slipped on the ice and broke her leg." *Why did she slip?* "Well, when you put ice under pressure, it melts. So momentarily Aunt Nellie was walking on a thin layer of water. Friction was reduced, so she slipped." *Why does ice melt when you put it under pressure?* "That's an interesting question! Most substances don't do that. The explanation eventually takes us into quantum mechanics." *Why did she fall* down, *not* up?

"Now that's a really deep question. . . ." Eventually, if we keep on asking "why" questions, we are led to the ideas of particle physics and then to the unanswered questions in particle physics.)

Every separate arrow of explanation points to a common source. Wherever we start, if we keep on asking "why" questions we will eventually end up discussing the basic building blocks of the Universe and the ways in which those building blocks can interact. It is in this sense that particle physics is fundamental. This is *not* to say that particle physics is more important, worthwhile, or interesting than, say, chemistry. But whereas we can hope to find an answer to a chemistry question in terms of fundamental physics, we would not expect to find an answer to a fundamental physics question in terms of chemistry. For example, "Why do we observe regularities in the periodic table of elements?" is a question that might be answered by physics. But the answer to "Why is the expansion of the Universe accelerating?" will not be found in chemistry.

Eventually, if we keep on asking "why" questions — and answering them with the aid of gedanken experiments — we might be led not to a "theory of everything," but to a universally applicable unified theory of space, time, the fundamental constituents of matter, and the interactions between them. We would then understand our Universe at a much deeper level than we do at present.

We will discuss various contemporary gedanken experiments later in this book. One of the most profound gedanken experiments, though, is over 2000 years old. The experiment was proposed by a philosopher rather than a physicist (physics had not been invented), but Plato's allegory of the cave can help us make sense of some of the most exciting physics ideas of the new millennium.

The Allegory of the Cave

To them, I said, the truth would be literally nothing but the shadows of the images.
Plato, *The Republic*

In one of his famous *Dialogues*, the great Greek philosopher Plato proposed the following gedanken experiment.

Imagine a group of prisoners chained in a cave in such a way that they are unable to turn their heads. All they can see — all they have ever seen throughout their conscious life — is the wall of the cave. Behind the prisoners, their captors maintain a fire; between the fire and the prisoners is a stage along which puppeteers can walk.

What do the prisoners see?

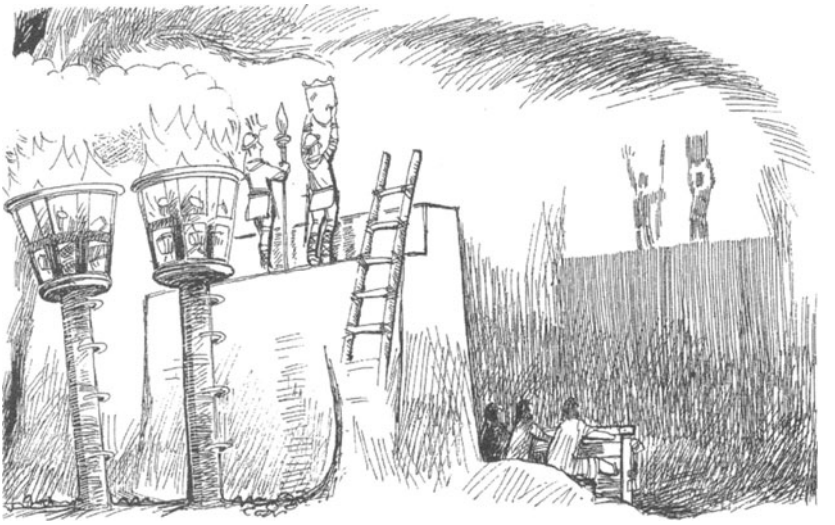


FIGURE 1 Plato's allegory of the cave. Prisoners are bound in such a way that they can see only the wall of a cave. A large fire causes shadows to be cast upon the cave wall. If a puppeteer holds up a shield, then a shadow of the shield is cast upon the wall. When the prisoners talk about a "shield" are they referring to the two-dimensional shadow, or to the real three-dimensional shield?

On the wall of the cave they see shadows cast by their own bodies. They also see shadows cast by the bodies of the puppeteers. And they see the shadows cast by any objects that the puppeteers hold up. In short, the prisoners are unable to see "real" objects; they only observe two-dimensional shadows cast by objects that they cannot see.

Since the prisoners could know nothing about the real causes of the shadows, they would think that the shadows were real. Plato argued that, under these circumstances, the prisoners would inevitably mistake appearance for reality. The prisoners would believe that they inhabited a two-dimensional world.

By carefully observing the shadows, any physicists among the prisoners might develop theories of how objects move and interact in their two-dimensional world. Such theories would necessarily be filled with inconsistencies, and the prisoners might wonder if they could ever find a "final" theory that explained the observed behavior of their world. It would be almost impossible for them to conceive of the notion that there is an underlying three-dimensional reality. Even if some of them broke free of their chains, looked out of the cave and glimpsed the intrinsic beauty of the three-dimensional world, they would find it difficult to persuade their fellow prisoners of the true nature of reality. Perhaps such people would be derided and mocked as madmen for having such wild ideas.

Although theoretical physics was not on Plato's agenda, his allegory may be astonishingly close to the situation that physics now finds itself in. As physicists pursue the fundamental laws, the picture of reality they work with is increasingly removed from everyday observations. Our theories of physics seem to tell us that we are in the position of the chained prisoners, bound to perceive only certain impressions when the world is actually much more interesting and beautiful than most of us imagine. Our theories imply, for example, that there may be more spatial dimensions than the three we perceive and that some these dimensions may be very much larger than previously thought. Some physicists suggest that our Universe may be just one of a number of parallel universes — and our Universe might have formed when two other universes collided. There is even a hint from modern physics that Plato's allegory may *literally* be true: perhaps our everyday reality can be represented completely by "shadows" on the walls of a higher-dimensional spacetime.

Plato's Allegory

In proposing the allegory of the cave Plato hoped to investigate some of the properties of his "realm of forms." He believed the notion of forms would address one of the deep mysteries of Greek philosophy: the question of change in the physical world. If, as fellow philosopher Heraclitus maintained, there is nothing certain except the fact that things change, then how can we know anything? How can we even *discuss* anything? How can we talk about dogs, for example, when dogs vary wildly amongst themselves and any individual dog — from the moment of its birth to the moment of its death — undergoes ceaseless change? Plato's answer to this conundrum appeared in one of the most influential books in Western philosophy. In *The Republic*, Plato argued that the *intelligible* world consists of the eternal *forms* of things. The form of a dog, for example, is abstract, unchanging, and applies to all dogs — and would continue to apply even if every dog in the world were to die. The intelligible world arises from human reason alone and is the world of reality. Humans live in the *visible* world — a world of change and uncertainty. However, the visible world is merely an imperfect manifestation of the perfect realm of forms. When we point at a dog and say "dog," we make sense only if we are referring to "dogness" — the form of a dog. It is this notion of form that Plato hoped to illuminate with his allegory of the cave. He of course had no inkling of the principles of modern physics — and the resemblance between his allegory and modern thinking on the nature of spacetime is purely coincidental.

These ideas are so far removed from our everyday experience that many people feel viscerally that the physicists must be wrong. But perhaps the physicists are the ones who have managed to look out of the cave, and the rest of us are prisoners condemned to look at shadows. Certainly we should not deride physicists for investigating these ideas, mock them as mad, or even edge away if one of them happens to sit beside us on a park bench. The purpose of this book is to try to make sense of these suggestions — to examine some of the wild ideas of modern physics.

The Search for Symmetry

There is only one good, namely knowledge,
and only one evil, namely ignorance.
Socrates

One of the difficulties we face when trying to understand the new picture of reality emerging from theoretical physicists (apart from the fiendishly difficult mathematics involved) is this: There is absolutely no experimental evidence for it. We have to rely upon gedanken experiments instead.

Physicists, in fact, are in an awkward situation. On the one hand, as we shall discuss in Chapter 2, they possess the two most successful theories in modern science. On a microscopic scale, the study of the fundamental constituents of matter and the interactions between them has led to the development of the so-called Standard Model of particle physics. (The Standard Model is discussed in detail in Chapters 3 and 4.) In principle, it provides a microscopic basis for *all* known physical phenomena (except, importantly, gravity). On a macroscopic scale, our best theory of gravity — general relativity — provides an explanation for phenomena ranging from pulse timings in the Global Positioning System through to the formation of the largest structures in the observable Universe. The success of these two theoretical edifices are the high points of twentieth century science.

On the other hand, there are good reasons for believing these theories are incomplete. The most important problem they face is that they are incompatible. At a fundamental level, general relativity and the Standard Model are in conflict. If we want a theory that unifies gravity with particle physics in a single, harmonious conceptual framework then we need some radically new ideas. For four hundred years, though, physics has been led by experiment. Where do we get new ideas, and how can we test their validity, if the outcomes of experiments are no longer guiding us? What are we to make of proposals that have no firm experimental justification?

The answer to such questions may lie in our very desire to possess a unified theory. We want a single conceptual framework to explain obser-

vations of physical phenomena because such a framework would be simpler, more beautiful, more *symmetrical* than a collection of mutually incompatible theories — even if both approaches explain the same of observations. We seek simplicity, beauty, *symmetry* in science just as much as we do in the arts. Indeed, symmetry is one of the oldest and most important unifying concepts in science. If experiment alone no longer shapes our theories, then perhaps aesthetics can help guide us. The search for new theories may become the search for new symmetries. The Greek philosophers, who sought symmetry in all its guises, would have appreciated this approach.

The search for symmetry has historically paid large dividends. Einstein was profoundly influenced by symmetry when he formulated his theories of relativity. Following his lead, theorists have looked wherever possible for symmetry in their theories, an approach which, when combined with results from a myriad of experiments, helped them develop the Standard Model. Interestingly, the same considerations of symmetry highlight the flaws in the Standard Model. Chapter 5 explains how we can alleviate the problems in the Standard Model by incorporating new symmetries — symmetries that have not yet been observed in nature.

If we make further leaps of faith — if, in addition to assuming new symmetries we also assume that there are other dimensions of space (see Chapter 6) and that the fundamental building blocks of nature are *not* the familiar point-like structures of particle physics (see Chapter 7) — well, then we can make progress toward a unified theory. The price we pay for such progress is a theory that is far removed from our everyday experience: a framework in which the fundamental objects are "branes" existing in a higher-dimensional spacetime and in which the inhabitants of one universe can be unaware that another universe exists within a millimeter of their own. These ideas are discussed in the Chapters 8–10.

This book, then, attempts to make sense of some of the wild ideas of modern physics by explaining the steps physicists have been forced to tread — forced by the demands of symmetry, though, rather than experiment. Ultimately, of course, we *must* get experiments to tell us if these ideas are merely a story, as much a tale as the allegory of the cave. Isaac Newton, the man who more than anyone else laid the foundations of physics, knew the importance of experiment. When told by a friend of observations that contradicted the Newtonian system (incorrect observations, as it turned out), he replied: "It may be so. There is no arguing against facts and experiment." The exciting possibility is that these wild ideas may soon be tested in our labs. In the next few years, experiments currently being planned may tell us whether we inhabit a Universe more strange and exotic than the one we observe — and that is the subject of the final chapter.

We begin the journey by looking at the crucial concept of symmetry.

1

SYMMETRY

Beauty is our weapon against nature; by it we make objects, giving them limit, symmetry, proportion. Beauty halts and freezes the melting flux of nature.
Camille Paglia, *Sexual Personae*

A polyhedron is the formal name for a solid bounded by plane polygons. This definition may seem rather abstract, but polyhedra are simple objects. They appear everywhere in our man-made, artificial world. An everyday example of a polyhedron is a breakfast cereal box: each face of the box is a rectangle (plane polygons with four edges), and the rectangles are joined together at the edges of the box so that it is closed. It is not difficult to see that we can construct a limitless number of different polyhedra. The *regular* polyhedra, though, are different.

A regular polyhedron has faces that are identical regular polygons — polygons with equal sides and equal angles, like the equilateral triangle and the square. A regular polyhedron is special because it possesses a high level of symmetry: it looks the same each time you turn it to present a new face. Contrast this with a breakfast cereal box, which presents different views as you turn it. (There is the large main face with the *Krispy Chocco Korn* name and logo on it, the smaller side face with the listing of ingredients, and the smallest top face that opens the box). The box thus lacks the symmetry of a regular polyhedron.

Remarkably, there are only five regular polyhedra: the tetrahedron, cube, octahedron, dodecahedron, and icosahedron (see Figure 2). No other regular polyhedra can exist. (The interior angles of the polygons meeting at a vertex of a polyhedron must add up to less than 360°. Only five regular polyhedra can satisfy this restriction.) This fact has long been known; in his book *Timaeus* (*c.* 350 BC), Plato described the five regular polyhedra, and in his honor these symmetrical objects are now called Platonic solids.

FIGURE 2 The five Platonic solids. From left to right: the tetrahedron (four equilateral triangles as faces), the cube (six squares), the octahedron (eight equilateral triangles), the dodecahedron (12 pentagons), and the icosahedron (20 equilateral triangles). In 2003, the mathematician Jeffrey Weeks singled out the dodecahedron as being worthy of special mention: his analysis of fluctuations in the cosmic microwave background provided intriguing hints that our Universe is in the shape of a dodecahedron.

One might well wonder why the Greeks bothered to study abstract constructs like the Platonic solids. The reason is simple: the Greeks were preoccupied with symmetry. They looked for symmetry in all areas of human endeavor, not just mathematics. They considered symmetry to be a unifying concept. This preoccupation has continued to a large degree in modern science. Symmetry is an important concept in biology, chemistry, crystallography, mathematics, and other fields. As we shall soon see, it also plays a basic role in physics. But what, exactly, *is* symmetry?

In everyday life, we have an intuitive notion of the meaning of symmetry: an object possesses symmetry if it appears to be the same object before and after we do something to it. For example, take a soccer ball, put your finger on the valve, and spin it: the ball looks the same while it rotates, and it looks the same from all directions. It possesses *rotational* symmetry. The cube — the second of the Platonic solids — also possesses rotational symmetry, but less than that of a soccer ball: only if we rotate a cube through an integer multiple of 90° — turning it through 90° or 180° or 270°, and so on — does it look the same as before the rotation took place.

The technical definition of symmetry is similar to our everyday understanding of the concept. A system has a symmetry if one can transform it in some way so that, after the transformation, it appears unchanged. In other words, it shows *invariance under transformation*. That is the essence of symmetry. The simplicity of the definition is deceptive; the search for theories that possess symmetry leads us to a new understanding of the physical world. Before considering the role that symmetry plays in physics, however, we can learn a lot by looking at some basic aspects of symmetry.

SYMMETRY IN EVERYDAY LIFE

Man is all symmetry,
Full of proportions, one limb to another,
And all to all the world besides.
George Herbert, *The Temple*

Consider the following transformation. Create a mirror image of a system by reflecting every point on one side of an axis into a point on the other side of the axis. If the system looks the same before and after this operation, then the system possesses *reflection* symmetry. This is perhaps the most recognizable of symmetries; it is what many people would point to if they had to give an example of symmetry.

The human body is bilaterally symmetric: reverse left and right, and it is difficult to appreciate that a transformation has taken place. The symmetry is not perfect, of course; the human heart is on the left, the liver on the right, and so the symmetry is broken. But to a good approximation our bodies possess reflection symmetry, and this symmetry is something to which all humans, regardless of culture, respond at a deep, almost visceral, level. For example, babies respond more positively to symmetrical shapes than to irregular ones; people generally regard a symmetrical face as more attractive than an asymmetrical face. (It is not clear whether our preference for symmetry has evolved because a symmetrical body indicates genetic fitness, or whether the human perceptual system processes symmetrical signals more efficiently than asymmetrical signals.)

FIGURE 3 This is a composite image, derived from combining images of several women. Studies consistently show that people regard such composite images, with their high level of reflection symmetry, as possessing beauty. Men and women of all cultures tend to rate a composite face as being more attractive than any of the individual faces that make up the composite.

Invariance under reflection (such as the bilateral symmetry of the human body) is rather different from invariance under rotation (such as the rotational symmetry of a soccer ball). The former is a *discrete* symmetry, the latter is a *continuous* symmetry. A continuous symmetry holds under an arbitrarily small transformation. For example, you can rotate a sphere through any angle — no matter how small — and the rotational invariance holds. This is not the case with a discrete symmetry, which has a "unit of transformation." For example, reflection in a mirror can be a symmetry transformation but one-third of a reflection is not. These two types of symmetry — continuous and discrete — have different mathematical properties but, as we shall see, both types play an important role in physics.

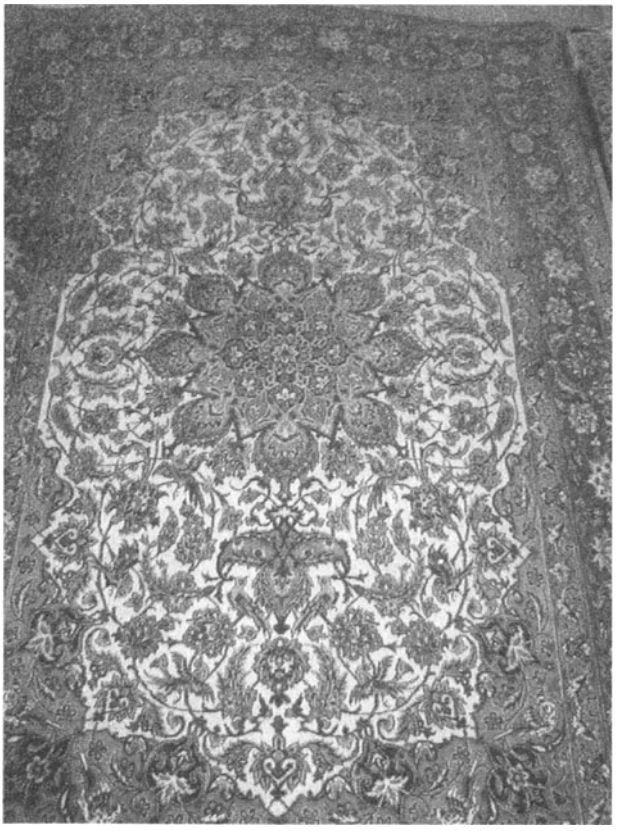


FIGURE 4 An Oriental rug like this may possess any one of a limitless number of motifs, but the repetition of the motif is constrained by symmetry. In fact, the repetition of a fundamental motif is subject to just four symmetry operations. The immensely skilled weavers repeat a motif by counting and then repeating a sequence of knots; they often deliberately break the symmetry — perhaps only slightly — in order to generate a design that is pleasant to the eye and intriguing to the mind.

Artists in all fields have played with the ideas of symmetry either by invoking symmetrical designs or by consciously breaking symmetrical motifs. Designs on Oriental rugs, for example, are often filled with various intricate symmetries. (There is usually a deliberate breaking of the symmetry, however. Variations add interest to an otherwise repetitive pattern.)

Symmetry is also a feature of patterned wallpaper. Patterns appear in wallpaper in such a way that, if you transform a piece of wallpaper by moving it a certain amount relative to its neighbors, the overall appearance is unchanged. Invariance under transformation: symmetry.

Symmetry and Wallpaper

You might think there is an infinite number of essentially different types of patterned wallpaper. You would be wrong. There are only 17 different types of wallpaper; all the patterned wallpapers you see in stores are merely variations on the 17 different themes. (This fact was first proved by Evgraf Fedorov, a crystallographer, in 1891. Fedorov was interested not in house decoration, but in the way in which repeating patterns can occur naturally in materials.) These wallpaper patterns were investigated in artistic fashion by the Moors of Spain. For theological reasons the Moors, who were strict Muslims, developed non-representational art forms to a high degree, many of which involved styles based upon geometrical forms. The Alhambra, a Spanish citadel which the Moors built in the thirteenth century, contains decorations that show *all* of the 17 different wallpaper patterns; Figure 5 shows three of the patterns.

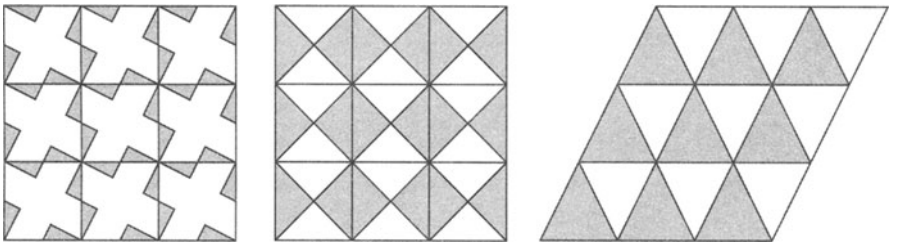


FIGURE 5 In 1891, the Russian crystallographer Evgraf Fedorov proved that there were only 17 fundamentally different types of pattern that could tile a plane. So there are only 17 fundamentally different types of patterned wallpaper. Here are three of them.

Buildings often possess symmetry, irrespective of the decorations they contain. Throughout the ages, architects have used symmetries in their plans. Many holy buildings possess reflection symmetry, perhaps because some cultures equate symmetry with perfection and thus only a symmetrical building is fit for God to dwell in. Secular buildings are also often symmetrical. Consider the Pentagon. It possesses rotational symmetry: if you could pick up the building and rotate it by 72° — one fifth of a full revolution — before placing it back on the ground, it would look the same as before

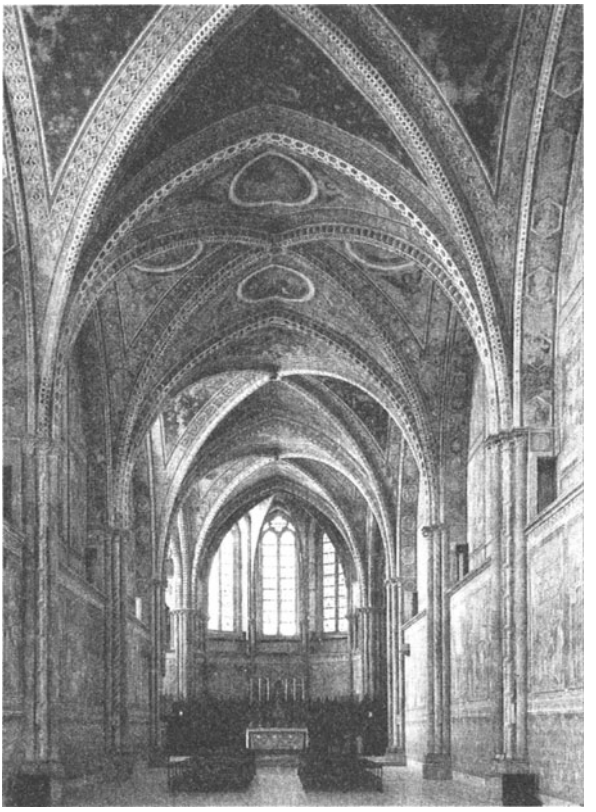


FIGURE 6 As in other art forms, symmetry is considered to be a unifying concept in architecture. Symmetry is particularly important in holy buildings; the Christian basilica — illustrated here by the Basilica of St. Francis in Assisi — always possesses bilateral symmetry. In these buildings, the axis of symmetry plays a symbolic role.

the transformation. This fivefold rotational symmetry is much less than the infinite rotational symmetry of a sphere, but it still makes for an impressive and recognizable structure. Some modern architects employ symmetry by deliberately breaking it — the Guggenheim Museum in Bilbao is a spectacular example.

Notions of symmetry — or deliberate breaking of symmetry — can be found in art of all types. Many of the paintings and woodcuts of the Dutch artist Maurits Escher, for example, depend upon the notions of symmetry; he often used the 17 wallpaper patterns to create playful effects. Some of Escher's work illustrates rather esoteric mathematical concepts; Figure 119 on page 239, for example, illustrates a negatively curved geometry using some visually striking symmetries.

Music is full of examples of symmetry — just think of Bach's *Musical Offering*. The rhyming and rhythm schemes of poems also often have an underlying symmetric structure. It is in nature, though, that we find some of the most beautiful examples of symmetry. In many ways the humble snowflake is the most striking.

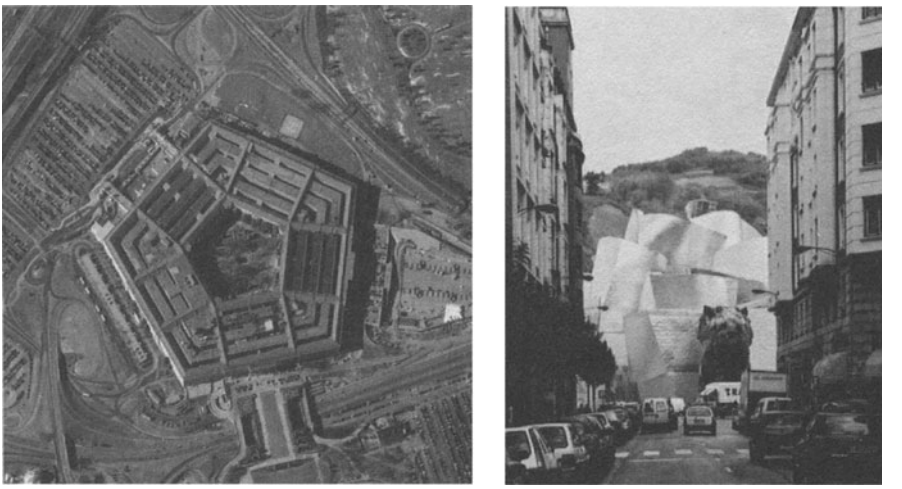


FIGURE 7 The Pentagon (left) possesses a fivefold rotational symmetry. In the absence of any distinguishing features on the five external walls, the building looks the same after a rotation through 72°. The Guggenheim Museum in Bilbao (right), on the other hand, has no obvious symmetries; we have so come to expect to see symmetry in buildings, that a structure like this can appear quite shocking.

SNOWFLAKES

How full of creative genius is the air in which these are generated!
I should hardly admire them more if real stars fell and lodged on my coat.
Henry David Thoreau, *Journal*

Snowflakes can be astoundingly intricate and spectacularly beautiful — and they all possess a simple sixfold rotational symmetry. How can a structure be so complex and yet be so simple? The answer to this question highlights a point that is of crucial importance for modern physics: the underlying laws of nature may possess simplicity and symmetry, but those laws can give rise to phenomena that are complex and difficult to understand.

The basic reason why snowflakes possess sixfold rotational symmetry has to do with the molecular properties of water and the way in which the two hydrogen atoms attach to the hydrogen atoms in other water molecules when water freezes. The molecules stack together to form a regular crystalline lattice, and the intramolecular forces involved ensure that the most stable configuration is hexagonal. The sixfold rotational symmetry of the underlying lattice means that all snowflakes have the same symmetry. There must be more to the story, however, since molecular forces act at the molecular scale, and a snowflake is many millions of times larger. How can such short-range forces generate symmetry on a large scale? The answer lies in

the way that an ice crystal can form certain smooth surfaces, called *facets*. The seed for an ice crystal — often a dust particle that has absorbed a few water molecules — will create an object with an irregular surface. These irregular surfaces have many "hooks" for water molecules that are floating about in the air, so such surfaces grow quickly into a plane. Facet planes, on the other hand, are much smoother on the molecular scale; they have fewer "hooks" with which to catch water molecules, and thus they grow much more slowly. The result is that the fast-growing rough surfaces quickly smooth out, leaving only slow-growing facet surfaces. Due to faceting, the seed quickly develops into a small hexagonal prism.

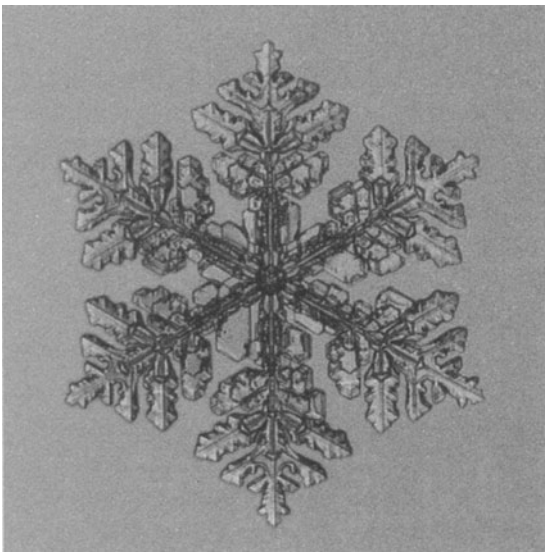


FIGURE 8 A snowflake, with its familiar sixfold rotational symmetry. The symmetry arises because, upon freezing, intramolecular forces cause water molecules to form a hexagonal lattice.

Even the small hexagonal prism of a new-born ice crystal is not the whole story. A small ice crystal looks rather like the head of a hexagonal bolt. How do we get from there to the intricate and enchanting structures with which we are familiar? Well, the corners of the crystal stick out into the surrounding air slightly more than the sides, and so the corners grow slightly faster and develop tiny tendrils. Since the atmospheric conditions are the same across the crystal (the crystal is tiny, after all — much smaller than the scale of changes in atmospheric conditions), the tendrils grow at the same rate, and the crystal maintains its sixfold symmetry. But atmospheric conditions can change; the temperature might decrease or increase, and winds might blow the crystal into regions of air with different temperatures. And, since the growth conditions are highly dependent upon atmospheric temperature, the crystal growth will vary wildly with time. It develops an intricate — and yet symmetric — shape. The result is beautiful

and unique: no two naturally formed snowflakes have the same shape, since no two snowflakes will have the same growth history. But all snowflakes possess the familiar sixfold rotational symmetry of the underlying hexagonal ice lattice.

At high temperatures, the elements that constitute a snowflake possess an even larger degree of symmetry: the water molecules whiz around equally in all three dimensions. The complex snowflake, with its lower degree of symmetry, freezes out of this simple set of initial conditions. What works for the snowflake may work for nature at large. As we shall see, the complex state of our present Universe may have frozen out of a much simpler state. Just as with the study of the snowflake, by studying the present symmetries we see in nature, we may be able to deduce the greater hidden symmetries of the Universe.

SYMMETRY IN PHYSICS

What immortal hand or eye,
Could frame thy fearful symmetry?
William Blake, *The Tiger*

The symmetry of snowflakes is just one type of symmetry found in science. As we progress through the book, we shall discuss many — often much more subtle — symmetries found in physics. The search for different symmetries has, indeed, been a driving force behind many recent theoretical developments. Before tackling the arcane symmetries of modern physics, however, it is worth looking at the central role that symmetry has played in well-established physics. A host of examples drawn from all branches of the subject could be given. The two examples I give here — Noether's theorem and the theory of special relativity — should be enough to demonstrate the power of symmetry in our quest to understand the Universe.

Noether's theorem

Some symmetries are so obvious, so intuitive, that it scarcely seems worth discussing them. It turns out, though, that an understanding of these symmetries helps us appreciate one of the most important pieces of knowledge we have about nature: Noether's theorem.

Consider the following transformation of a physical object: move (or, to use the jargon, *translate*) it from one place to another. Such an operation is called a *spatial translation*, and it is about as simple as a transformation can be. A moment's thought tells us that this is, in fact, an exact symmetry

transformation. Pick up a chair from your dining room, move it into your living room, and then try to think of any physical property of the chair that has changed because of the translation. The subatomic particles that constitute the chair do not change after it has been moved; the molecular properties of the chair, which give rise to qualities such as color and strength and shape, do not alter after the translation. Therefore any equation with which physicists might care to describe the chair — an equation for its load-bearing capacity, perhaps, its electrical resistance, or its light-absorption properties — must be invariant under translation in space. The same comments apply to any system and any spatial translation; the laws of physics look the same before and after a transformation that consists of moving an object in space. Invariance under transformation: that is the hallmark of symmetry. Nature's laws are symmetrical under spatial translations. Our Universe did not *have* to be that way: we can imagine a universe with a crystalline structure, for example, in which the outcome of an experiment *did* depend upon the direction of translation. But our Universe is not like that.

Here is another example of a transformation: a *temporal translation* — the translation of a physical object in time. (Such a transformation does not require the use of a time machine; it simply means that any measurement we make on a system can be made at different times — now, in the past, or in the future.) It turns out that the laws of physics are invariant under temporal translations. If you were to measure the electrical resistance of your dining room chair, for example, then you would observe the same result today as you observed yesterday and as you will observe tomorrow (assuming that all other conditions are the same, and that you have not, and will not, alter the chair in some way). Nature is thus symmetrical with respect to temporal as well as spatial translations.

Yet another simple example is the symmetry of the laws of physics under rotations. For example, if your dining room chair is facing north and then you turn it so that it faces east, you will find that the properties of the chair remain the same.

These three symmetries — invariance of the laws of physics under spatial translation, temporal translation, and rotation — are observational facts. One day, someone may observe a phenomenon that breaks one of these symmetries, but that day has yet to arrive. Physicists, astronomers and geologists have performed many experiments and recorded a myriad of observations, and the symmetries have never been observed to be broken. Any equations that purport to describe the Universe must therefore be invariant under these three simple transformations. Why, though, should this seemingly trivial fact be so important?

An invariance under transformation means that something is unchanged — or *conserved* — through the transformation. So it is not surprising that there is a connection between symmetry and the most basic, comprehensive, and powerful laws of physics that we possess: the conservation laws. In 1915, the German mathematician Emmy Noether proved a famous theorem. For every continuous symmetry of the laws of physics there exists a conservation law, and for every conservation law there exists a continuous symmetry. This theorem is one of the reasons why physicists search so assiduously for symmetries.

FIGURE 9 Amelie (Emmy) Noether

Consider a collection of objects — perfectly elastic pool balls, perhaps — rattling around inside a closed, isolated container. We can translate the container and its contents through space, and the physics inside the container is unchanged. From Noether's theorem, the symmetry of physics under spatial translations implies that linear momentum is conserved. So whenever and however the pool balls inside the container interact, their total linear momentum before and after the interaction must be the same. This is the familiar law of conservation of linear momentum, which is demonstrated every day on pool tables around the world. Conversely, the myriad of experiments — from pool games to particle accelerators — that have demonstrated the accuracy of the conservation of linear momentum have also proved that space is symmetric under spatial transformations. Space is homogeneous.

The symmetry of physics under temporal translations implies, via Noether's theorem, that energy is conserved. However the objects in a closed system interact their total energy before and after the interaction must be the same. This is the law of conservation of energy. And all those effects proving the accuracy of the conservation of energy — effects taking place every day in power stations, for example — also prove the laws of physics are the same now as they were in the past and as they will be in the future.

Finally, the symmetry of physics under rotations implies, via Noether's theorem, that angular momentum is conserved. Whenever and however the objects in a closed system interact, their total angular momentum before the interaction is always the same as their total angular momentum

afterward. This is the law of conservation of angular momentum. Again, arguing in reverse, all those observations that have confirmed the conservation of angular momentum — observations ranging from the way ice skaters spin faster when they tuck their arms close in to their bodies through to the way neutron stars change their rotation rate following a starquake — have also demonstrated that space is symmetric under rotational transformations. Space is isotropic.

We take the conservation laws for granted, but our familiarity with them should not blind us to the fact that they are remarkable and fundamental pieces of knowledge about our Universe. And ultimately they derive from some simple symmetry principles.

You might think we have ignored one of the great invariances of the Universe in the discussion above: the invariance of physics between an interchange of left and right. This is *mirror symmetry*. If you perform some experiment on your dining room chair, surely you get the same results if you perform the same experiment on a mirror image of the chair.

The general term for the oddness or evenness of something is *parity*. In the case of a mirror transformation, the term is *space parity*. The space parity of an object characterizes how that object behaves when the coordinates that describe it are reversed. Until 1956, physicists assumed that physics must be unchanged under a space parity transformation: looking at an experiment in a mirror should not affect its outcome. Just as the homogeneity of empty space leads to the conservation of linear momentum and the isotropy of empty space leads to the conservation of angular momentum, you might think that the reflection symmetry of empty space leads to the conservation of space parity. Not so!

It turns out that space parity is *not* conserved; empty space *does* distinguish between left and right. Certain fundamental particles are inherently left-handed in some interactions. They possess helicity in the same way that a left-handed screw has a particular thread direction. Upon reflection in a mirror, helicity is interchanged and a left-handed screw becomes a right-handed screw. So nature looks different when viewed in a mirror.

When this fact was discovered in the late 1950s, it shocked many physicists. Nowadays, they take the fact for granted. Our fundamental theories of physics must give rise to certain processes that do not possess mirror symmetry. In the jargon of the field, nature is *chiral*. (The word "chiral" comes from the Greek word for hand; in other words, nature has an inherent handedness.) Wolfgang Pauli, one of the most brilliant physicists of the twentieth-century, put it best: "God is left-handed."

So finding the symmetries of a theory is important. Finding the symmetries that a theory does *not* possess — or the ways in which the symmetry of a theory is broken — can be equally important.

Special relativity

Einstein's theory of special relativity is grounded in notions of symmetry and invariance. (Einstein actually wanted to call it "invariance theory.") An appreciation of the particular symmetry of special relativity gives us another fundamental piece of knowledge about our Universe.

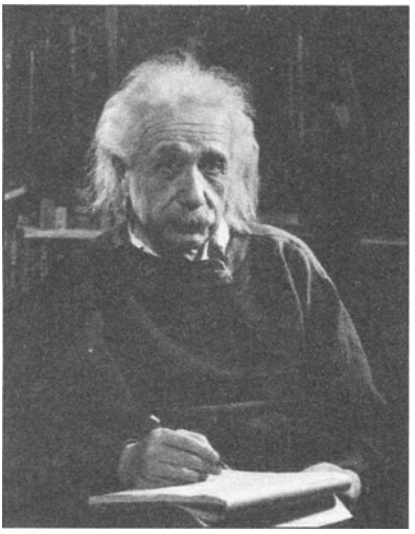


FIGURE 10 Albert Einstein

Einstein published his theory of special relativity in 1905. Before then, scientists used Galilean physics when they calculated quantities such as length and speed. For example, suppose you need to measure the length of your dining room chair. To do this you would associate a set of three coordinates with the bottom of the chair leg — one coordinate for each dimension of space. (As we shall argue in later chapters, there may be more spatial dimensions. But certainly we observe only three spatial dimensions, and back in the seventeenth century, René Descartes showed that we need three numbers — coordinates — to identify a point in three-dimensional space.) Then you would associate three more coordinates with the top of the backrest. Then you would calculate the distance between these points using a simple mathematical recipe that dates back to Pythagoras: the square of the distance between the two points is equal to the sum of the squares of the differences between each of the coordinates. In practice, of course, you would simply take a tape measure and read off the length — but doing that is equivalent to the long-winded procedure I have just described.

Now suppose you ask your friend to measure the length of the same chair using the Pythagorean method outlined above, and to do so he turns the chair so that it lays flat on the ground. He would have a completely different coordinate system and would associate different sets of coordinates with the endpoints. But the equations of physics obey rotational symmetry, so the length of the chair must be the same under the two sets of measurements; your friend would find exactly the same answer for the square of the distance between the two points. In other words, this quantity — the square of the distance between two points — is invariant under rotations, although the individual coordinates change because of the rotation. You can even use this observation as a *definition* of rotational symmetry.

Notice there is no mention in the above procedure of *when* the measurements take place. In Galilean physics, space and time coordinates are quite separate; all *inertial observers* (observers who move freely and with no external forces acting on them) measure the same time. Time is universal.

Einstein disagreed. He extended the concept of invariance under rotation to take account of time. He began by analyzing a thought experiment in which inertial observers are in relative motion and considered what such observers would measure. His deliberations led him to point out that our physical world consists of a network of *events* — phenomena that occur at an instant in time and at a point in space. When we do physics we observe and measure events. Events must be labeled with a set of four coordinates: three to specify the point in space and one to specify the instant in time. *But the time coordinate is not independent of the space coordinate.* Hermann Minkowski, Einstein's teacher, later concluded that space and time coordinates are so deeply entwined that we must talk about four-dimensional *spacetime* as a single entity.

This viewpoint is at odds with our naive understanding of how the Universe works. According to our mental models, we live in three-dimensional space, and time is something quite different from space. Our mental models are pragmatic: they evolved in the way they did because, in our environment, it is entirely reasonable to take the duration of objects for granted. Objects generally change and move slowly, so time can be factored out. But our mental models are wrong. Einstein was right, as countless experiments have shown. We live not in space and time, but in spacetime.

A point particle, as it moves through spacetime, traces out a line — its *world line* — which is the entire history of the particle. (Every particle in your body traces out a world line, and if some godlike observer could step outside spacetime it would see your whole life as a collection of particle world lines; they would converge at the moment of your birth, wriggle like a worm through spacetime, then go their separate ways when you die.) By considering the world lines of photons and how they connect events, Einstein was led to the notion of the *spacetime interval.* The spacetime interval between two events is analogous to the spatial distance between two points. All inertial observers measure the same value for the spacetime interval between two events. For example, consider the following events: a firecracker explodes at some point in space and another firecracker explodes at a different point in space. Just as you and your friend measure the same length for your dining room chair even if your views of the chair are rotated, so will you measure the same spacetime interval between the two explosions even though you may be in constant relative motion. *Any* inertial observer, no matter how she is moving, would measure the same spacetime interval between the two explosions.

The transformation that changes an observer's velocity is called a *boost*. So Einstein was arguing that the spacetime interval is invariant under boosts. Invariance under transformation: there must be a symmetry involved. There is indeed a symmetry, now called *Lorentz invariance* after the Dutch physicist Hendrik Lorentz who first arrived at some of the key ideas of relativity. (Noether's theorem tells us that Lorentz invariance must have an associated conservation law. There is indeed a conserved quantity associated with Lorentz invariance, but it has only technical significance and we will not consider it further.)

Einstein argued that the fundamental laws of physics must possess the symmetry of Lorentz invariance. If they did *not* have this symmetry, different observers would measure different fundamental physics purely because they were in relative motion. It would be like different observers measuring different lengths for a chair simply because the chair had been rotated.

So special relativity says that any equations purporting to describe the Universe must contain the symmetry of Lorentz invariance. This is a simple requirement, yet some seemingly bizarre phenomena emerge from it. Any good book on relativity will explain the details behind relativistic phenomena (several are given in the bibliography, and the reader is urged to consult them). What can be lost in many textbook treatments, however, is an appreciation that these phenomena emerge from the demands of symmetry.

The formula for the spacetime interval is similar to the Pythagorean formula for the distance between two points. However, we measure distance and time in different units. To treat them in a unified way, we must introduce a conversion factor between space and time; in other words, we must introduce a fixed speed. This fixed conversion factor, which is given the symbol c, is the speed of light. But c is much more important than the particular speed at which photons happen to travel. It is a basic physical constant, which calibrates the relationship between time and space. The fundamental point here is that there exists a speed such that the spacetime interval is invariant under rotations and boosts.

The symmetry of Lorentz invariance is really all there is to special relativity. This symmetry, however, gives rise to effects that are not usually observed in our low-speed world. For example, the spacetime interval between events connected by a light signal is always zero. This means every observer measures the same result for the speed of light, regardless of how they are moving. If Galilean physics were correct, you could chase after a light beam at a speed $3c/4$ and you would observe the light beam to have a speed $c/4$; if you traveled at speed c alongside the beam the observed light speed would be zero; if you traveled faster than light you could overtake the beam. All of this is impossible according to special relativity, because all observers measure the same speed, c, for light.

Relativistic Addition of Velocities

If all observers measure the same speed for light, it follows that the familiar rules for adding velocities must be wrong. For example, if a ray gun can shoot bullets that travel at $0.9c$, and you fire it forward on a rocket moving through space at $0.9c$, then the speed of the bullet to an Earthbound observer would *not* be $1.8c$. Instead, it would, according to the rules of special relativity, be $0.9945c$. This is an indication of one of the famous predictions of relativity: no material object can travel faster than the speed of light.

Scientists before Einstein did not realize that a new formula was required for the relativistic addition of velocities, because the velocities they dealt with were so small compared to the speed of light. It was only in 1905, the year that Einstein published his theory, that an automobile first exceeded 100 mph. But 100 mph is just 0.0005 percent of the speed of light — no wonder relativistic effects could be ignored!

Before Einstein, the relationship between the coordinates measured by different inertial observers was given by a set of formulae known as the *Galilean transformation*. The Galilean transformation describes our everyday world of low speeds with great precision; it says that we all measure the same time, that the lengths we measure do not depend upon velocities, and so on. It accords precisely with our everyday experience. But we know the Galilean transformation is incorrect, because it does not possess Lorentz invariance. The correct set of formulae that relate the coordinates measured by different inertial observers — the *only* simple formulae that preserve Lorentz invariance — is known as the *Lorentz transformation*. And the Lorentz transformation mixes up the time and space coordinates as Einstein told us it must. Inertial observers who travel at relative speeds that are close to the speed of light will disagree about the times measured by their clocks; this is the phenomenon of *time dilation*, which is often described by the phrase "moving clocks go slow." Inertial observers in relative motion will also disagree about the lengths they measure; this is the phenomenon of *length contraction*. As we move faster our clocks tick slower and our rulers become shorter. (So an astronaut travelling in a spaceship at near-light speed relative to Earth would measure smaller lengths for objects than an Earthbound observer. The astronaut would also age more slowly. Upon returning home, he could find that his friends had grown old and died while he remained young.)

Most quantities that physicists measure in their laboratories are measured, ultimately, using clocks and rulers. But if clocks can slow down and

rulers can contract, it follows that we need to look again at laboratory-defined quantities that we might think of as constant. In particular, Einstein pointed out that energy is defined in terms of measurements involving timepieces and yardsticks. We thus need to consider our definition of energy with care. For example, a bullet shot from a gun has a certain energy. Suppose we could build a "super-gun" that could shoot the bullet at speeds close to light speed; the symmetry of Lorentz invariance would cause us to measure different values for many of the bullet's properties. Einstein showed that the measured mass of the bullet would increase as the bullet moved faster, and this increase in mass would come from an increase in the bullet's energy. Einstein was thus led to the most famous prediction of special relativity: the equivalence of mass and energy, through the equation $E = mc^2$. The equation says that we can convert energy into mass and vice versa. Because the factor c^2 is so large, a small speck of matter can produce a vast amount of energy, as was evidenced in the explosions of atom bombs. This was a stunning result. Previously, physicists had thought that energy and mass were separately conserved; now, there was the single law of the conservation of mass–energy. After Einstein, it was clear that matter could be created or destroyed — so long as there was a corresponding change in energy. Mass was just another form of energy — a very condensed form, to be sure, but energy nonetheless.

Time and space, matter and energy are not the absolute, invariant quantities that everyday life tells us they are. They are very much observer-dependent quantities. They *have* to be, because that is the only way of ensuring that a deeper symmetry holds true: all inertial observers measure the same spacetime interval between events. The seemingly bizarre phenomena of special relativity occur because spacetime possesses symmetry — the symmetry of Lorentz invariance.

GROUPS: THE MATHEMATICS OF SYMMETRY

Mathematics is the study of pure pattern,
and everything in the cosmos is a kind of pattern.
Rudy Rucker, *Mind Tools*

The language of physics is mathematics. The particular dialect of mathematics that describes symmetry is called *group theory*. Symmetry is such an important concept that the terminology of group theory appears frequently in physics literature. Although this book is intended for a general audience and so contains very little mathematics, a brief description of some of the most common terms of group theory may be useful.

Symmetries of a Snowflake

We saw that snowflakes can exhibit almost infinite variety, but that their symmetry properties are limited: there are relatively few transformations under which snowflakes are invariant. Indeed, if you represent the snowflake by a regular hexagon, then it is a straightforward matter to list the transformations that leave it unchanged. First, you can rotate a regular hexagon through 0°, 60°, 120°, 180°, 240°, and 300°, and the sides of the hexagon return to their original position. There is a sixfold rotational symmetry. Second, you can reflect a regular hexagon in six different mirror planes and find that the hexagon appears unchanged. If you play around with a couple of cardboard cut-outs of regular hexagons you will quickly convince yourself that these are the only transformations that leave a hexagon unchanged. The hexagon has a symmetry of order 12. Snowflakes may possess infinite variety, but they possess only 12 symmetry transformations. Symmetry has thus brought some order to the bewildering complexity of snowflakes. (Figure 11 illustrates these symmetries.)

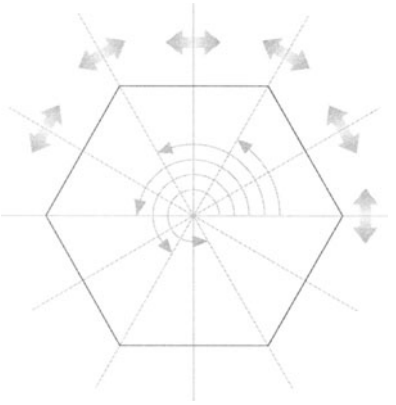


FIGURE 11 Symmetries of a hexagon

The set of symmetries of the hexagon possesses four properties. First, the set is *closed*. (Closure means if we perform one symmetry transformation on the hexagon and then immediately follow it by another symmetry transformation, the result is one of the set of 12 symmetry transformations.) Second, the set is *associative*. (Take three transformations, called A, B, and C. Do A followed by B then follow it by C. Note what happens, then start again. This time do B followed by C; then do A followed by the result of B followed by C. The result should be the same as the first procedure. This sounds complicated, but it essentially means that the order in which transformations take place does not matter.) Third, the set has an *identity*. (An identity is a transformation that does nothing. In this case it is a rotation about 0°.) Fourth, every transformation in the set has an *inverse*. (Each transformation can be followed by another transformation that returns the hexagon to its original state, as if the identity transformation had acted upon it.)

Any set of objects that possesses closure, associativity, identity, and inverse is known as a group.

To a mathematician, a group is the technical term for any set of entities that possesses four simple properties (the boxed text on page 26 describes these properties). Mathematicians have studied groups for more than 170 years — ever since Évariste Galois, a young Frenchman, wrote down some key results in the subject. Group theory has since generated a *vast* literature. It has become so much a part of mathematics that the basics of the subject are often taught to schoolchildren. The more advanced techniques of group theory are also tremendously powerful, and a wide range of problems can be solved where no other approach would succeed. Fundamentally, though, groups are simply a description of symmetry. So whenever scientists uncover a symmetry — whether it is a symmetry of crystals, molecular vibrations, or the fundamental constituents of matter — they need only identify the appropriate group and then they have an "off-the-peg" set of mathematical results and tools with which to analyze the symmetry.

In the case of the symmetries of the snowflake, the appropriate group is called the dihedral group. It does not matter what we *call* the group, however; the important point is that its properties have been studied and cataloged by generations of mathematicians and scientists. So, if a physicist has determined that a sample of material possesses hexagonal symmetry, she can use the well-established results pertaining to the dihedral group in her study of the various properties of that material.

There are many different types of group, each type corresponding to one of the many different varieties of symmetry transformation that exist. A discrete symmetry, such as the reflection symmetry of a snowflake, is the easiest type of symmetry to visualize; and the discrete symmetries certainly play an important in fundamental physics. The more abstract continuous symmetries, though, and the groups that describe them, are even more important. The so-called *Lie groups*, in particular, play a crucial role. They seem to provide a description of the symmetry properties of the most basic constituents of the Universe.

FIGURE 12 Sophus Lie

The Lie groups are named after the nineteenth century Norwegian mathematician Sophus Lie (pronounced "Lee"). Lie managed to catalog the groups of continuous symmetries into several classes, and since his pioneering work, mathematicians have classified all possible continuous symmetries. (The complete classification of discrete symmetries was a much more difficult task and has only recently been completed).

An important class of Lie group is called SO(N), or the special orthogonal group of order N, where N can be any integer. The name itself is unimportant. Stripped of the jargon, the group SO(N) simply describes rotational symmetry in N-dimensional space. Even that may sound complicated until we realize that an object with this type of symmetry could be a smooth sphere: an object like a basketball, for example. Pick any axis around which to rotate a basketball: no matter which angle you rotate it through in three dimensions, the ball rotates back into itself. It is invariant under rotations. And this rotational symmetry of the basketball is described by the group SO(3) — where the "3" stands for the three dimensions of space.

Lie also described groups known as U(N): *unitary* groups of order N. The simplest example of this type of group is U(1), and we shall see how the equations that describe electromagnetism possess this symmetry. Lie also described groups known as SU(N): *special* unitary groups of order N. These are abstract symmetries, so it is less easy to point to an object and describe its symmetry in terms of SU(N). The essence of these symmetries is simple, though: the SU(2) symmetry can "rotate" two objects into each other and the SU(3) symmetry can "rotate" three objects into each other. We shall see later how these symmetries play an important role in describing fundamental physics.

The groups SO(N) and SU(N), along with another class of Lie group called Sp(N), the symplectic group, are known as the "classical" groups. They have had a long history of applications in physics. In addition to the classical Lie groups, there are five *exceptional* Lie groups: G(2), F(4), E(6), E(7), and E(8). These groups are much more tightly constrained than the classical groups. The symmetries of the exceptional groups are not symmetries that one can demonstrate with everyday objects like snowflakes or basketballs. However, mathematicians have deduced their properties through abstract reasoning. The largest and most mysterious of the exceptional groups is E(8), which is a 248-dimensional object whose properties mathematicians are still investigating. As we shall see, the exceptional group E(8) seems to play a role in theories that attempt to unify gravity with the other forces.

Just as knowledge of the dihedral group is useful for scientists studying ice crystals, so knowledge of groups like U(1), SU(2), SU(3) and E(8) is useful for scientists studying the fundamental constituents of matter. The literature on the mathematical structure of these objects is intimidating. Fortunately, we only need to bear one thing in mind: all these groups simply describe different types of symmetry.

2

TWIN PILLARS OF PHYSICS

Einstein's special and general theories of relativity have permanently changed our view of space and time and gravitation. In an even more radical break with the past, quantum mechanics has transformed the very language we use to describe nature: in place of particles with definite positions and velocities, we have learned to speak of wavefunctions and probabilities. . . . But now we are stuck.

Steven Weinberg, *Dreams of a Final Theory*

Many undergraduate physics courses contain an element known as "modern physics." What is usually meant by "modern physics" in this context is two different topics: relativity and quantum mechanics. These two theories are undoubtedly the twin pillars of modern physics, the foundations upon which our contemporary understanding of the Universe is built. And they feel modern, somehow, in that they deal with concepts that are quite unfamiliar to us in everyday life. Relativity, for example, tells us that our measure of time depends upon our position in a gravitational field; quantum mechanics forces us to give up the notion

that an object simultaneously possesses a definite position and momentum. These concepts appear to be, in that horrible phrase, bleeding-edge concepts. Yet it is worth remembering that these "modern" ideas now have a long history, in contrast to many of the speculative ideas we will meet later in this book. The theory of special relativity is almost a century old; Einstein's extension of those ideas, which forms the theory of general relativity, was published about nine decades ago; quantum mechanics was cast into its present form not long after that. So these "modern" ideas are neither untried nor untested; quite the contrary — they have met with spectacular success, which is why we teach them to our students.

A theory's success is gauged by the extent to which the numerical predictions of the theory agree with observations. If there is close agreement between the values we calculate and the values we measure, the theory is viewed as successful. By this definition both general relativity and quantum mechanics work extremely well. This is comforting from a purely intellectual viewpoint, but there are practical benefits from possessing successful theories. An example of such a benefit is the development of the Global Positioning System (GPS), a technology that requires precise applications of both relativity *and* quantum mechanics for its operation. I begin this chapter with a discussion of GPS, since its very existence highlights a paradox at the heart of physics: both general relativity and quantum mechanics are in some senses "correct," but they are fundamentally incompatible.

The Global Positioning System

I once was lost, but now am found.
John Newton, *Amazing Grace*

The Global Positioning System gets my vote as the most amazing of recent technologies. Millions of people, from yachtsmen and automobile drivers to hikers and climbers, now rely upon GPS radio signals to tell them exactly where they are on Earth's surface. The military, of course, makes heavy use of GPS, but the system is also used to monitor the movement of tectonic plates, to track the movements of elephants in Africa, to locate faults in power lines, and to time stamp financial transactions. Each year a host of new applications of the system are launched.

The system has 24 satellites orbiting 20,000 kilometers above Earth. These are positioned in such a way that, from most places on Earth, at least four of the satellites are above the horizon at any time. Since each of the satellites orbits Earth in exactly half a day, any fixed observer on the ground will see any particular satellite at exactly the same place in the sky twice

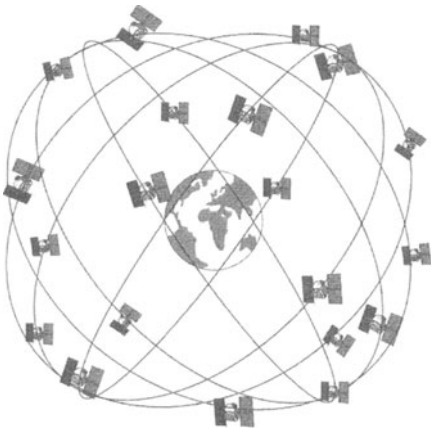


FIGURE 13 The Global Positioning System uses 24 satellites orbiting 20,000 kilometers above Earth. From most places on Earth's surface, at least four of the satellites are above the horizon at any time.

each day. Each satellite carries an atomic clock. The phenomenal accuracy with which atomic clocks measure time is combined with a fundamental principle of special relativity — namely, that the speed of light is a constant for all observers — to make a precision navigational system. The system works by having the satellites continually broadcast coded radio signals that can be detected by GPS receivers (which nowadays can be small enough to wear on the wrist). Suppose a receiver on Earth's surface simultaneously gets a coded signal pulse from each of the four satellites above its horizon. Since the signals carry information on when the pulse was transmitted and it is known that the signals travel at light speed, a computer inside the receiver can calculate the distance to each of the satellites. And these four pieces of information give enough data to calculate the precise coordinates of the receiver; it is essentially just a high-tech form of triangulation — the process whereby a particular point can be located by viewing it from two different, accurately known points — which has been used by navigators for centuries.

The procedure outlined above may seem straightforward, but there are some intricacies to account for before engineers can construct a working navigation system.

Suppose we require GPS to provide a positional accuracy of one meter. That means the atomic clocks on board the satellites have to be accurate to about one part in 10^{13}. No problem: atomic clocks possess this order of accuracy, and in any case the clocks are frequently corrected by a reference timekeeper at the US Naval Observatory. However, as we shall soon see, relativistic effects are much larger than one part in 10^{13}. In fact, they are about ten thousand times too large to ignore. The intricacies of relativity — of Einstein's view of spacetime — must be addressed in any Global Positioning System.

FIGURE 14 GPS is a US Defense-Department system designed to provide worldwide, three-dimensional positioning information. Civil customers may also access the system, although at reduced accuracy. Researchers in the 1970s used the first generation Block I satellites to test the feasibility of the GPS concept. The present system uses the second generation Block II satellites like the one shown here.

We saw earlier how the symmetry of special relativity demands that inertial observers in relative motion measure different times between the same events. In rough language, "moving clocks go slow." Well, the atomic clocks on board a GPS satellite are certainly moving relative to a receiver on Earth's surface. Typically, a satellite speed is about 4 kilometers per second, which causes the atomic clocks to be slow by about one part in 10^{10} — roughly one second in 300 years. Furthermore, since the orbits of the satellites are elliptical, the satellites speed up and slow down as they change their distance from Earth. (This is the result of the conservation of angular momentum, which as we know is another symmetry effect.) So the relativistic time dilation varies throughout the orbit; all this has to be calculated and taken into account by the receiver. But that's not all. General relativity predicts the existence of phenomena that cause even larger effects on the timekeeping of GPS clocks. As we shall see later, clocks in different parts of a gravitational field may register different times. Well, the atomic clocks on board a GPS satellite are high up in Earth's gravitational field: 20,000 kilometers higher than the receivers. A basic calculation shows that a clock on board a GPS satellite runs *faster* than an identical ground-based clock by about five parts in 10^{10}. Again, the size of this "gravitational time dilation" varies throughout the orbit as the satellites change their orbital distance from Earth. Another general relativistic phenomenon — the so-called Sagnac effect — arises because, relative to fixed receivers on Earth, the satellites are rotating. All these effects — and other, even more subtle, relativistic effects — must be accounted for if GPS is to provide positional accuracy of one meter.

FIGURE 15 A GPS receiver must be small enough to be practical for hikers and climbers; a receiver such as this one is no larger than a cell phone. A GPS device like this continually receives coded radio signals from GPS satellites. At any given instant, signals will be received from four of the 24 satellites. A computer inside the device calculates the distance to each of the four satellites. Since the positions of the satellites are known with accuracy and precision, and the coded signals carry information about when signals were transmitted, the location of the GPS receiver can be determined to within a few meters.

It is often stated that general relativity has been tested in only a few extreme astrophysical situations; that it has little or no practical use in the everyday world. This is simply wrong. If relativistic corrections are not applied to the time registered by the atomic clocks on board GPS satellites, the system does not work. The theory of general relativity plays a key role in one of the most exciting of modern technologies.

Implicit in the above discussion is an engineering capacity that we often take for granted. Our ability to build atomic clocks; to construct computers that are powerful enough to handle the difficult mathematics involved in calculating relativistic corrections and yet are small enough to fit inside a portable receiver; to develop the communication networks required for synchronizing the system: these are now so much a part of our everyday world that they scarcely seem worthy of remark. Yet this capability, which is based on our ability to manipulate matter and fields on a microscopic scale, is almost magical. The gadgets of Western civilization — televisions, computers, smart washing machines, MRI scanners, and the rest — all work because physicists have learned how to control the flow of electrons. And this has come about because of our grasp of another pillar of modern physics: quantum mechanics. Without a knowledge of quantum mechanics, it would be impossible to develop a technology like the Global Positioning System.

But there is a paradox here. In the Global Positioning System we possess a well-established technology that simultaneously employs both general relativity and quantum mechanics. Without general relativity we would get the timings incorrect, and the system would be worthless; without quantum mechanics, we could not build the system in the first place. The system requires both quantum mechanics and general relativity to be in some sense "correct." *Yet the fundamental assumptions behind the two theories are incompatible.*

Most physicists are now as certain as they can be about anything that the Universe is governed by quantum laws. But it turns out that when quantum ideas are applied to general relativity the result is nonsensical. How can we have two incompatible fundamental theories of physics? Would it not be more elegant, more beautiful, more symmetrical, to have *one* theory with which to explain physical phenomena?

The drive to eliminate this incompatibility — the attempt to make general relativity coexist with quantum mechanics — has led to the "wild ideas" that we discuss later: strange new symmetries, extra dimensions, and parallel universes. Before we can appreciate these ideas, we need to look a little more closely at the twin pillars of physics in order to understand the source of the incompatibility between them. Let's begin with a discussion of general relativity.

Physics on a Large Scale: The Theory of General Relativity

Isn't gravity brilliant?
Brilliant, *The Fast Show*

The theory of *special* relativity embedded the beautiful symmetry of Lorentz invariance into physics, but Einstein was not satisfied. His theory, or rather the rest of physics, was not yet symmetrical enough. His pursuit of symmetry in physics — his quest to write down theories that ensured the same laws of physics apply to everyone, regardless of where they are or how they move — led him to the theory of *general* relativity. General relativity is our best theory of gravity and of how the Universe works on a large scale. Let's look briefly at how Einstein came by his theory.

Special relativity identified inertial observers as being important; it was they who agreed upon the invariance of the spacetime interval between two events. But what about non-inertial observers? What about those observers who underwent acceleration or who felt a force? In particular,

what about those observers who were in the grasp of a gravitational field? How could relativistic notions be applied to them?

Einstein developed his theory of general relativity from a simple gedanken experiment; it led to a realization that he called "the happiest thought" in his life. Imagine that you are in a small, sealed, windowless room. If the room is not accelerating, then you cannot tell whether or how fast you are moving; the principles of special relativity ensure this. Indeed, the notion of assigning a velocity for the room makes no sense unless you can look outside and measure your motion relative to some external object. You can consider your room to be at rest. Things are very different if an acceleration is involved. This time there is a force, pressing you to the floor; you *feel* it, regardless of whether you can see outside. We call such forces "gee forces." The constant acceleration due to gravity on Earth's surface is 1 *g*, but on fairground rides we can briefly encounter larger forces.

Now, suppose there is a scale in the room. You stand on it, and it registers your normal weight. Is there any way of telling whether the room is on the surface of the Earth, or is in a spaceship that is accelerating at a constant 1 *g*? Einstein thought about this simple question for years, and concluded that the answer is "no." There is thus an intimate relationship between gravity and acceleration. Einstein realized that a local gravitational field is indistinguishable from an accelerated motion — he called this the *equivalence principle*. It is another example of a beautiful symmetry in physics.

If there is no difference between an accelerated observer in the *absence* of a gravitational field and a non-accelerated observer *with* a gravitational field, then we can always assume that the latter perspective is valid. We can say that *all* observers, no matter what their state of motion, can consider themselves to be at rest, as long as they include a suitable gravitational field to describe their immediate environment. Thus Einstein generalized his ideas of relativity.

The equivalence principle leads directly to a surprising conclusion involving space. Consider again the sealed, windowless room that is undergoing an acceleration; perhaps the room is in an accelerating rocket ship. Suppose that you stand at one wall and shine a laser beam horizontally at the opposite wall. In the time it takes for the light ray from the laser to traverse the distance to the wall, the ship will have accelerated beneath the ray. The light ray will hit a point on the opposite wall that is slightly lower, relative to the floor, than it set out. In other words, the acceleration causes the ray to follow a curved path. But the equivalence principle tells us that a gravitational field must also bend the path of a light ray. So why is this surprising? Well, a light ray always takes the path that requires the least time between two points; this is a well-observed and established principle of physics. Usually, we think that the path with the smallest time between

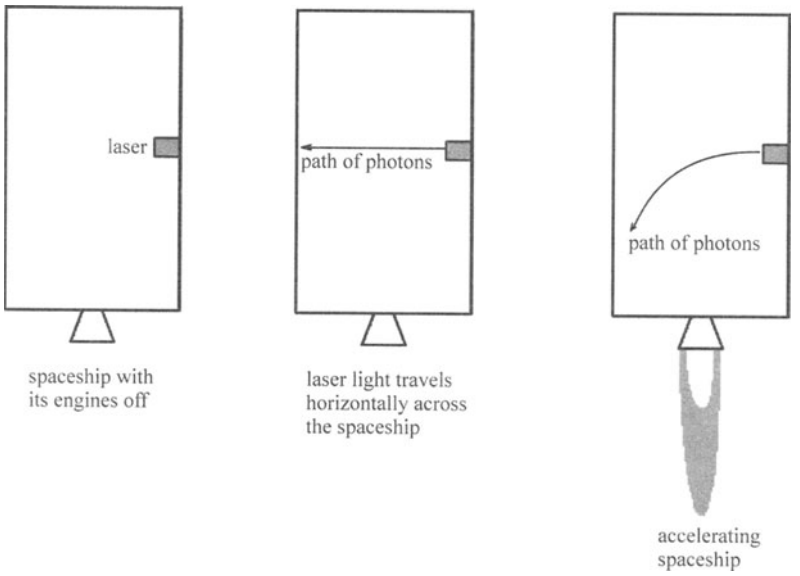


FIGURE 16 The equivalence principle tells us that photons "fall" in the presence of a gravitational field.

two points is a straight line, and light rays travel through space in straight lines. We therefore usually think of space as being like some flat, Euclidean plane. (Light rays can be refracted, of course; light rays are bent when they hit glass after travelling through air, for example, which is a phenomenon that enables us to build spectacles and telescopes. Even when light is refracted like this, the light rays follow the path that takes the least time. Here, though, we are considering the simplest case of light travelling through a vacuum.) If light rays bend in a gravitational field, then the path with the shortest time between two points must be a curved line, not a straight line. Thus where a gravitational field exists, *space is curved!*

The equivalence principle leads to another surprising conclusion, in this case involving time. We can understand it best by employing a similar simple gedanken experiment.

Imagine that our rocket ship is in space and has been using its motors to keep a constant position relative to Earth. The captain then cuts the motors. Instantaneously, the ship is stationary with respect to Earth; then it starts to accelerate under the influence of Earth's gravitational field. The ship is in free fall, and people on the ship are weightless.

A laser, fixed to the ceiling of a room in the ship, points vertically down and it fires a pulse at the instant the motors are cut. If you were directly beside the laser when the pulse was emitted, you would measure a particular frequency for the light. This measured frequency would be the

"standard" frequency for the laser — the value that would appear in the manuals and the reference books. An observer on the room's floor would measure the same frequency, because the conditions in the ship are the same as those in an inertial frame with no fields present. But what about an observer who is standing on Earth's surface, and is thus stationary in Earth's gravitational field? What frequency would he measure for the laser beam? He would agree with you about the "standard" frequency, since the laser was at rest relative to Earth when the light left the ceiling. But by the time the light hits the floor, the ship will have accelerated in Earth's gravitational field, and will be moving downwards with some velocity. So the observer who is stationary relative to Earth's gravitational field will measure a higher frequency — in much the same way that an observer hears a higher pitch from a siren as an ambulance speeds towards him. (The phenomenon is known as the Doppler shift. Radiation emitted by an object moving toward an observer is "squeezed;" its frequency appears to increase and is therefore said to be *blueshifted.* Radiation emitted by an object moving away from an observer is "stretched;" its frequency appears to decrease and is therefore said to be *redshifted.*)

In other words, if light shines downwards in a gravitational field it is shifted to a higher frequency (blueshifted). If light shines upwards in a gravitational field it is shifted to a lower frequency (redshifted). The generic name for this phenomenon is *gravitational redshift,* and it is a direct result of the equivalence principle.

The surprising conclusion arises when we recognize that the atomic oscillations that emit the laser light are, in effect, extremely accurate clocks. A clock is, after all, merely a device that repeats an action at a well-defined and constant rate. The action can be mechanical, like the ticking of a watch, or something more rarefied like the atomic oscillations that give rise to the emission of laser light. It follows that *time passes at different rates at different places in a gravitational field.* Time passes more slowly in a strong gravitational field than it does in a weak gravitational field. (This effect is clearly measurable. The US atomic standard clock, which is accurate to 1 microsecond per year, is kept at Boulder; an identical clock is kept at the Royal Greenwich Observatory in England. Boulder is about 1650 meters above sea level, whereas Greenwich is roughly at sea level. The clock at Boulder gains 5 microseconds per year over the clock at Greenwich. Clearly, GPS must take gravitational redshift into account.)

So not only does a gravitational field warp space, it also warps time. More elegantly, since Einstein showed how space and time are inextricably linked, we can say that a gravitational field warps spacetime. In fact, Einstein argued that gravity *is* a warp, a distortion, in the geometry of spacetime!

This was a fascinating insight by Einstein, and it causes us to review the way that we look at the world. It means, for instance, that Earth does not orbit the Sun because it is impelled to do so by some force. Rather, Earth simply follows the straightest path it can in spacetime, but spacetime is curved due to the presence of the Sun's mass and energy.

The notion that spacetime can be curved follows directly from the equivalence principle. But Einstein went beyond this. His theory of general relativity provided a detailed description of gravity. The equations of general relativity gave a precise recipe for calculating the curvature of spacetime caused by a given distribution of mass and energy; similarly, it allows one to calculate the trajectory of a particle through a curved spacetime. (The equations of general relativity are beautifully simple in appearance, but often fiendishly difficult to solve. The famous quotation by John Wheeler captures the essence of the equations: "Matter tells space how to curve; space tells matter how to move.") The theory generates many more predictions than gravitational redshift and spacetime curvature. It predicts the existence of black holes and gravitational waves; it accounts for the orbits of celestial bodies; it provides us with our best framework for understanding cosmology — the study of the Universe as a whole.

As well as being successful, general relativity is widely regarded as being the most beautiful theory in all of physics. One of the reasons is that, in developing the theory, Einstein introduced a very deep symmetry to physics: *general coordinate invariance*. In other words, the theory looks exactly the same no matter how an observer chooses her coordinates of space and time. By combining this deep symmetry and the equivalence principle with beautiful mathematical concepts in geometry, Einstein's theory flows naturally.

According to Newton, there is a gravitational force between any two massive objects that is proportional to their masses and inversely proportional to the square of the distance between their centers. This formula is accurate enough for almost all practical purposes. Using Newton's ideas about gravity and motion we can land spacecraft on Mars; we need only Newtonian ideas when we calculate the trajectories of projectiles. But there are cases — like the refinements necessary for a functioning Global Positioning System — when only the full theory of general relativity will do. The physics of the macroworld is the physics of general relativity. But general relativity tells us nothing about the microworld — a regime where other fundamental forces are more important than gravity. The microworld is best described by quantum mechanics.

PHYSICS ON A SMALL SCALE: THE WEIRD WORLD OF THE QUANTUM

I think it is safe to say that no one understands quantum mechanics. Do not keep saying to yourself, if you can possibly avoid it, "But how can it be like that?" because you will go into a blind alley from which nobody has yet escaped. Nobody knows how it can be like that.

Richard Philips Feynman, *The Character of Physical Law*

The microworld contains entities like electrons, which we call *particles*. In classical physics, the word "particle" conjures up an image of something hard and featureless — something like a tiny billiard ball, perhaps. In classical physics, you can apply a label to a particle like a billiard ball; and then, if you know the forces acting on the ball and the interactions it undergoes, you can follow the motion of the particle exactly. When, in the early twentieth century, physicists tried to apply these ideas to describe the behavior of entities like electrons it became clear that quantum particles do not obey the rules of classical mechanics. Laws that work superbly for describing the motion of tiny billiard balls fail completely to describe the motion of electrons. An electron is *not* a scaled-down version of a billiard ball, nor of any other familiar everyday object. It is something altogether more weird.

Depending upon the experimental setup you use to investigate the properties of an electron, you might see an electron behave like a particle in some ways. (It might trigger a detector at a precise location, for example.) It might also behave in some ways like a wave. (It might diffract off a crystal, for example.) The basic entities of nature simultaneously possess aspects that in classical physics are thought to be separate; the basic entities exhibit *wave–particle duality*.

Consider electromagnetic waves, which were once thought to be described completely by Maxwell's classical theory of electromagnetism. There are indeed situations in which electromagnetism can be described as a wave phenomenon, but there are also situations in which electromagnetism is best described in terms of particles. When objects emit electromagnetic radiation, they actually emit a quantum particle — called a *photon* — which has wavelike properties such as frequency and wavelength.

Wave–particle duality is found whenever we probe the behavior of matter on small scales. The unified description of this dual behavior is provided by quantum mechanics.

So we now have two systems of mechanics: one for the microworld (quantum mechanics) and one for the macroworld (classical mechanics). At the very least, it is unaesthetic to have two different systems of mechanics. Can we not reduce this to a single theory, a single system of mechanics?

It turns out that we can.

The scale at which quantum phenomena become noticeable is given by a fundamental constant of nature, called the *Planck constant* (symbol h). The Planck constant governs all quantum phenomena. For example, an electron moving along with some momentum has an associated wavelike property with a wavelength given by h divided by the momentum. It is this wavelike property that is missing in classical mechanics. Or consider a stream of photons, which behave like a wave with a frequency and a wavelength. Each photon has an associated particle-like property with an energy given by h multiplied by the frequency and a momentum given by h divided by the wavelength. It is this particle-like property that is missing in classical mechanics. As h tends to zero we recover our familiar everyday classical mechanics. Classical mechanics is thus a limiting case of quantum mechanics, and it is valid when the Planck constant can be neglected. *There is thus only one system of mechanics: quantum mechanics!*

The Planck Constant

The value of the Planck constant, h, is about 6.6×10^{-34} J s. It has the dimensions of *action* (energy multiplied by time).

The combination of h divided by 2π occurs so frequently in physics that it is given its own symbol: $\hbar$, pronounced "aitch-bar."

Now, the Planck constant has a small value. In everyday life we *always* deal with quantities that dwarf the Planck constant, and it is therefore always valid to neglect it in our usual activities. Because the Planck constant is so small, we never see quantum effects as we go about our daily business. Space, time, and energy always appear continuous; particles behave as particles, and waves behave as waves.

But although h is small, it is not zero. When we investigate the microworld we inevitably deal with quantities that are comparable in magnitude with the Planck constant. It is in these situations that quantum mechanics is inescapable.

Perhaps the phenomenon that best encapsulates the difference between classical and quantum mechanics is the *uncertainty principle*. To appreciate this aspect of quantum behavior, consider what we have to do in order to "see" an electron and establish its position. To view any object we have to bounce photons off it. But we have to use photons with a wavelength less than the size of the object, otherwise the photons will bend around it without scattering from it; in other words, if their wavelength is too long, the photons are not scattered back into our eyes or microscopes. So to

Quantum Mechanics in Macroscopic Systems

Quantum mechanics certainly rules at the level of individual particles, where the effects of the uncertainty relations make themselves felt in a myriad of ways. Increasingly, though, quantum effects are being observed in much larger systems.

Recently, for example, Robert Knobel and Andrew Cleland at the University of California, Santa Barbara, have observed the uncertainty principle at work in what is almost a macroscopic system. They worked with a tiny, vibrating crystal bar (just a micron in length and cooled to very low temperatures) and a motion detector that could record displacements as small as one percent of the diameter of a single atom. All this might seem like a microscopic system, but the crystal bar had a mass that was the equivalent of 10 billion atoms. With their setup, Knobel and Cleland saw the crystal bar — an essentially macroscopic system — vibrate due to effects arising from the uncertainty principle.

Or take another example of a purely quantum phenomenon that has now been observed in a macroscopic system: entanglement. Entanglement occurs when the quantum state of two or more objects must be described with reference to each other, even if the individual objects are separated by large distances. Entanglement leads to correlations between physical properties of the objects that are stronger than any correlations allowed by classical physics. If two particles are entangled, measuring the state of one particle enables us to infer the state of the second particle without touching it. In the next few decades, entanglement will allow the development of wonderful new technologies — such as quantum teleportation and quantum computing — but most studies of entanglement to date have involved subatomic particles. In 2003, however, Thomas Rosenbaum, at the University of Chicago, and his colleagues in Britain and America, observed the phenomenon of entanglement in a single crystal of magnetic salt. This is a tiny system compared to our everyday world, but compared to subatomic scales it is macroscopic. We simply cannot escape quantum mechanics.

"see" an electron and establish its position precisely, we require *very* short-wavelength light. However, from the wave–particle duality relationships given above, a very small wavelength implies a very large momentum for the photon. Some of this momentum will be imparted to the electron when the photon hits it. We can establish the position of an electron with precision, but only at the expense of accepting an uncertainty in the electron's momentum. Alternatively we can use longer-wavelength light. This

will cause less disturbance to the electron — but we trade increased precision in our knowledge of the electron's momentum with decreased precision in our knowledge of its position. We cannot measure both quantities simultaneously with complete certainty. Indeed, as numerous experiments have shown, an electron does not simultaneously *possess* a well-defined position and momentum. In quantum mechanics, then, the product of the uncertainties in the pair of position/momentum measurements has an irreducible value. (This value is given by $\hbar$; here again, we see how the Planck constant governs the microworld.) This is the uncertainty principle.

There is a related uncertainty relation involving time and energy. We cannot measure precisely the energy of a particle over very small time intervals. On small enough timescales, the energy of a particle — even the energy of empty space — can fluctuate wildly.

The uncertainty principle makes it seem as if we must abandon all hope of having a mathematically rigorous theory of quantum particles such as the electron; it seems that we can have no real idea of where an electron will be nor what velocity it will possess. As luck would have it, things are not that bad. In the 1920s, the Austrian physicist Erwin Schrödinger argued that although the electron is a point particle, it is always accompanied by a wave. The wave is a function of the particle's position and certain other variables. This electron wavefunction obeys a mathematically rigorous and well-defined equation — the Schrödinger equation. The wavefunction tells us everything there is to know about the particle. The kick, though, is this: the square of the wavefunction gives us only the *probability* of finding a particular particle in a particular state. (The wave is some sort of probability wave.)

The wavefunction is a purely quantum idea, and it gives rise to behavior that is impossible in classical physics. For example, we can confine an electron to a box. If the electron has insufficient energy to jump over the walls of the box then, according to classical physics, the electron will remain inside the box for ever. But according to the rules of quantum mechanics, the electron wavefunction can seep outside the box. This means there is a probability — perhaps a small probability, but a non-zero probability nonetheless — of finding the electron outside the box. According to quantum mechanics, an electron can *tunnel* out of a box. This phenomenon of tunneling is employed in a quantum-mechanical device called a tunnel diode, which is used in amplifiers for some satellite communication systems. Without the purely quantum phenomenon of tunneling, such diodes would not work — and we would not have a technology like the Global Positioning System.

There is thus a randomness lying at the heart of physics. The uncertainty principle imposes a limit on what we can know; if we try to calculate

"the future" from "the present," we get only statistical results, a probability distribution. For any given process we can calculate only what is *likely* to happen, not what *will* happen.

Quantum mechanics is an essential tool in many different branches of science, including astrophysics, chemistry and crystallography. In many of these cases the velocities of the systems being studied are slow enough that we can ignore the effects of special relativity, and non-relativistic quantum mechanics can be used. What concerns us most in this book, however, is the fundamental particles and their interactions. When we probe the nature of fundamental particles, the velocities involved are often comparable to the speed of light and special relativity cannot be ignored. If we hope to understand the nature of fundamental particles travelling at near-light speed, we must combine quantum mechanics with special relativity. Ensuring that this combination worked correctly was not easy and required two decades of effort from some of the world's most brilliant minds. The combination was pretty much in its final form by about 1950: it became known as *relativistic quantum field theory*, and it provides us with our best description of the fundamental particles. It provides a conceptual framework for understanding many of the truly fundamental aspects of particles and the interactions between them. Although the topic contains some difficult technicalities, the basics can be appreciated without recourse to mathematics. In Chapters 3 and 4 we look at relativistic quantum field theory in a little more detail; here, though, it is worth looking at what the words mean. We have already met the terms "relativistic" and "quantum," but what is a "field?"

The field concept is an old one, and it was employed to great effect by classical physicists. An example of a classical field is ocean temperature. At every point in the ocean we can assign a number that corresponds to the temperature of the water at that point. Ocean temperature is a *scalar* field: a *single* number gives the magnitude at each point in space. The number varies from point to point (different parts of the ocean are at different temperatures) and it varies with time (the ocean gets hotter or colder depending upon the season and the time of day).

A *vector* field associates both a number and a direction with each point in space. The velocity of water throughout the ocean is a vector field: at every point in the ocean we can associate a number — the speed of the water at that point — plus the direction in which the water is moving.

Fields allow disturbances to move within them. We usually call these disturbances waves. Mathematically, these disturbances are described by *field equations*. Thus there are field equations for the propagation of the scalar temperature field and the vector velocity field of water in the ocean. In

this case the field equations are complicated, because the fields can interact — the scalar temperature field can affect the vector velocity field (since a temperature differential between points will tend to cause water to move) and vice versa. In a complicated situation like this, computers must be used to solve the equations. But the *form* of the equations was known to classical physicists in the nineteenth century.

One feature of the two examples given above is that the temperature and velocity fields require a medium (in this case, the ocean) in which to propagate. Other examples of fields, such as air pressure and string tension, also require a medium in which to propagate (air and string, respectively). These fields can only be defined in terms of the properties of some other physical entity. But what about the electromagnetic field? In what medium does the electromagnetic field propagate? Einstein, in developing his theory of special relativity, showed that electromagnetic fields require no medium in which to propagate. They *can* propagate in a medium — light travels through glass, after all — but a physical medium is not *necessary* for them to propagate. The electromagnetic field is in some sense fundamental, or elementary: we cannot define it in terms of other things. The same is true of certain other fields we shall meet in the next two chapters. We should not seek to explain such fields in terms of the properties of other entities. The fundamental fields are the simplest things we know about. They are irreducible — they have no "parts," they are not made out of other "stuff." Instead, when we keep on asking "why" questions in the manner of Richard Feynman and Steven Weinberg, we should seek to explain all other entities in terms of the dynamics of the fundamental fields. The study of basic physics is the study of the fundamental elementary fields.

James Clerk Maxwell introduced the electromagnetic field many years before physicists appreciated the quantum nature of the Universe. (Maxwell's theory is thus a classical theory.) But in the 1930s and 1940s, physicists learned how to make the electromagnetic field obey the dictates of quantum theory. The same recipes could be used to quantize other fundamental fields. When the fields were also made to obey Lorentz symmetry — when quantum mechanics was married with special relativity, in other words — the result was relativistic quantum field theory.

Quantum field theory (the "relativistic" is implied) provides us with some important insights. For example, associated with each fundamental field is a quantum particle: *the fields manifest themselves in nature as particles.* So the electron is the quantum particle manifestation of something called the Dirac field, the photon is the quantum particle manifestation of the electromagnetic field, and so on. A startling conclusion of quantum field theory is that the number of particles need not be conserved. This means, for example, that photons and electrons can be created "from nothing," so

long as there is sufficient energy for the process to occur and the various conservation laws are obeyed. This phenomenon is routinely observed in particle physics experiments. Particle creation is simply not part of classical physics; we need quantum field theory to explain it.

Another striking conclusion, first reached in 1928 by the British physicist Paul Dirac, is that each of the fundamental quantum particles must have a corresponding *antiparticle*. A particle and its antiparticle have exactly the same mass, but they possess "mirror-reversed" properties. The negatively charged electron, for example, has the positively charged positron as an antiparticle. (The photon, which has zero charge and zero mass, serves as its own antiparticle.) When a quantum particle interacts with its antiparticle, the two mutually annihilate. For example, if an electron reacts with a positron they annihilate and are replaced by photons. The amount of energy liberated in such an annihilation process can be tremendous, since the factor c^2 in Einstein's equation $E = mc^2$ is so large. (Although antimatter is routinely created at particle accelerators, albeit in minute quantities, the world around us consists entirely of matter, with no antimatter to be seen. Since the laws of physics do not obviously favor the existence of particles over antiparticles, the absence of antimatter in the Universe is one of the deep mysteries faced by physicists and astronomers.)

Yet another remarkable result from quantum field theory is this: the single Dirac field can describe any number of electrons — which means that *all* electrons are *exactly* identical. In the quantum world, if you have seen one electron, you have seen them all. The same is true of other fundamental particles. This is quite different to the classical situation, were we can "label" particles and thus distinguish between them. This leads to quantum phenomena that are quite inexplicable in terms of classical ideas.

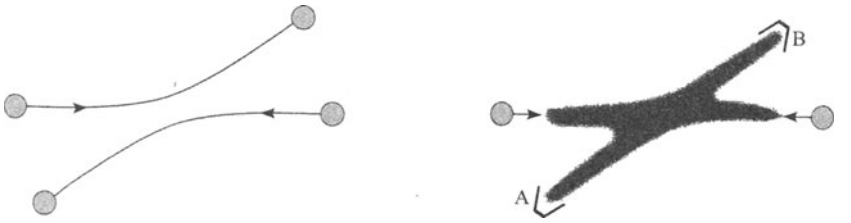


FIGURE 17 Quantum particles differ from classical ones in several ways. For example, identical quantum particles are exactly *alike. This gives rise to several interesting phenomena, one of which is shown here. The left-hand diagram shows the collision between two classical particles; we can attach a label to these particles so we can follow their trajectories throughout the collision. The right-hand diagram shows the collision between two quantum particles. The uncertainty principle causes a "fuzziness" in their trajectories. When the two particles come close together, their wavefunctions overlap and it becomes impossible for us to distinguish the two particles. We have no way of knowing which particle is detected at A and which at B.*

All of this will be discussed in more detail later. What is important here are the following two facts. First, the Universe is inherently quantum in nature; we fail to notice this in our everyday life only because the Planck constant is so tiny. At small distances, we inevitably encounter quantum mechanics. Second, the most successful theories in all of science — the ones that pass the test of experiment better than any other — are relativistic quantum field theories.

When GR Meets QM: Big Trouble at Little Distances

God does not play dice with the cosmos.
Albert Einstein

The two pillars of physics — general relativity and quantum mechanics — have quite different realms of applicability. We use general relativity to describe the behavior of objects on a large scale; we use quantum mechanics to describe the behavior of objects on a small scale. We might hope that, since the theories apply to different situations, we can live with two fundamental theories. That hope is in vain.

It is messy from an aesthetic point of view to have two quite different theories of nature, but it is much more than an aesthetic problem. If some of the fundamental physics interactions are classical, then we could use only those interactions in order to obtain the precise position and momentum of a particle — thus violating the uncertainty principle. If *some* of the fundamental laws of physics are quantum, then they *all* have to be quantum. Besides, there are some situations — close to the singularity of a black hole, for example, or in the immediate aftermath of the Big Bang — where we need *both* theories if we want to understand what is going on. These are situations where spacetime is very highly curved on very small distance scales. On both practical and aesthetic grounds we require just a single fundamental physical theory. So what happens when we try to combine general relativity with the principles of quantum mechanics? What happens when we try to write down a quantum field theory of gravity? Unfortunately, the result is nonsense.

General relativity is a classical field theory. So is electromagnetism, for that matter, and physicists know how to quantize the electromagnetic field. By following certain procedures physicists have obtained a spectacularly successful relativistic quantum field theory of electromagnetism. As in all other walks of life, the prime rule in physics is: if something works,

copy it. So physicists followed the same procedure in order to quantize the gravitational field, and indeed they can write down a quantum theory of gravity. There is a quantum particle — the *graviton* — associated with the gravitational field, just as the photon is the quantum particle associated with the electromagnetic field. In this picture, gravitational processes occur through the exchange of gravitons, just as electromagnetic processes take place through the exchange of photons. One can calculate the result of physical processes involving the graviton, just as one can with the photon. But the results of these calculations have an awkward tendency to be infinite. The result of any calculation in a quantum theory is a probability — the chance of a particular event occurring. The result, in other words, lies between 0 (no chance of the event happening) and 1 (the event is certain to happen). What are we to make of a probability that is infinite? In fact, we can make nothing of it. The appearance of infinities is a sure sign that the theory is wrong. For more than five decades, some of the brightest minds on the planet have attacked the problem of infinities in quantum gravity. So far, no one has managed to banish the infinities. They seem endemic in this kind of approach.

Why does the approach to quantization that works so well for our classical theory of electromagnetism fail so badly for our classical theory of gravity? The main reason is that the central assumption of general relativity is that spacetime has a smooth geometry. On large distance scales this assumption works just fine. However, as we probe microscopic distance scales the assumption crashes into the uncertainty principle and breaks down. The uncertainty principle tells us that quantum fluctuations cause undulations in the gravitational field — even in the zero gravitational field of an empty universe. As we probe smaller and smaller volumes of space, the uncertainty principle tells us that the quantum fluctuations become bigger and bigger. At the *Planck length* of 10^{-35} meters, space itself becomes a roiling, raging tumult. Our familiar notions of spacetime — the concept of "here" and "there," of "then" and "now" — begin to lose their validity. In these conditions, at the Planck length, general relativity fails because the assumption on which it is founded (namely a smooth spatial geometry) fails. The equations of general relativity no longer hold at the Planck length, and this breakdown of the theory manifests itself in infinities.

Even before Einstein, physicists tried to unify gravity with the other forces. But progress was slow. In the absence of progress, most physicists chose to investigate *either* the microworld of quantum mechanics *or* the macroworld of general relativity and left unification to some later date. This strategy paid off handsomely, leading us to a much deeper understanding of nature. At last, though, there are signs — just tentative glimpses — that unification

might happen. But to appreciate these developments, we need, like the physicists, to consider the two pillars of physics separately. We continue the story in the next chapter with a closer look at the microworld — the world that is so precisely described by relativistic quantum field theory.

3

BUILDING BLOCKS OF THE UNIVERSE

It is the most persistent and greatest adventure in human history, this search to understand the universe, how it works, and where it came from.
Murray Gell-Mann

Although our world is a wonderfully complex place, people have long suspected that this complexity arises from the interactions of a rather limited number of building blocks. The Greeks, for example, believed that there were four elements — fire, earth, air, and water — from which all else was made. Nowadays, we know that there are 92 naturally

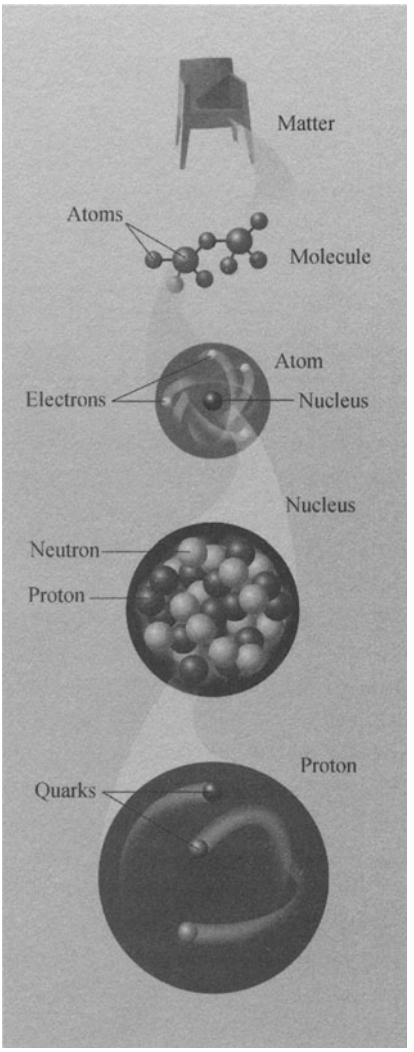


FIGURE 18 Trying to discover the ultimate building blocks of matter is rather like unpeeling the layers of an onion. We know that matter in the everyday world — the material that makes up a chair, for example — consists of molecules. The molecules are made up of atoms. Atoms possess an internal structure: they consist of a cloud of electrons orbiting a nucleus. The nucleus itself possesses an internal structure: it consists of protons and neutrons. And the protons and neutrons are also composite objects: they consist of three quarks. At the present time, it seems that the electrons and the quarks are point-like, structureless objects. They are believed to be fundamental particles. But many theorists argue that there is good reason to believe that these particles are themselves made out of something even more fundamental — strings.

occurring chemical elements; they range from the lightest element, hydrogen, to the heaviest element, uranium. All the material that we see around us in our everyday world consists of one or more of those 92 elements in varying combinations. Eventually, scientists discovered that elements differ because the atoms that constitute them are different; a hydrogen atom is different from a uranium atom. The source of this difference became clear once scientists realized that atoms, rather than being indivisible, possess a structure. An atom consists of one or more electrons orbiting a central nucleus, and the atomic nucleus itself possesses structure: a nucleus consists of one or more protons and a variable number of neutrons. (The property

that identifies an atom is the number of protons in its nucleus; different isotopes of an element possess the same number of protons in the atomic nucleus, but differing numbers of neutrons. The simplest hydrogen atom consists of one electron orbiting one proton; the most common isotope of uranium has atoms formed of 92 electrons orbiting a nucleus consisting of 92 protons and 146 neutrons.) More recently, scientists discovered that protons and neutrons themselves possess a structure; they consist of three quarks bound tightly together.

Elements, atoms, nuclei, protons, quarks: whenever we investigate the fundamental structure of matter, we seem always to find another layer. There are indications, though, that we may have uncovered the true, fundamental particles: the building blocks from which all matter in the Universe is constructed. Before looking at why we might have hit bedrock, let's first look at how we produce the fundamental particles.

Producing the Fundamental Particles

O amazement of things — even the least particle!
Walt Whitman, *Song at Sunset*

The electron and the photon are the most familiar of the truly fundamental particles. The familiarity stems in large part from the ease with which we can produce them. Electrons are a key constituent of atoms, so ordinary matter is chockful of them. And we need relatively little energy to liberate them; simply heat a piece of metal, and electrons come boiling off. Once we have produced electrons, we can then use electromagnetic fields to manipulate them. The production of photons is even easier: the energies involved in lighting a match are sufficient to produce photons in abundance. Unfortunately we require much higher energies if we wish to study the other fundamental particles. The problem is not so much the energy, in fact, but the need to pack that energy in a small enough region to probe structures that are much smaller than the size of a proton.

Cosmic rays — particles from space that constantly bombard Earth's atmosphere — pack a large punch in a small package. The most energetic of these naturally occuring objects possess far more energy than we can ever hope to produce by artificial means, so studying the debris after a cosmic ray collides with an atom in Earth's atmosphere is a good way to look for fundamental particles. (Figure 19 shows a typical cosmic ray detector.) Unfortunately we have no way of knowing when or where a cosmic ray will strike, what its composition will be, or how much energy it will possess. So although cosmic ray studies are valuable — and we shall see in Chapter 11

just how much — it is difficult to control the observations. The best way to produce particles in a controlled manner and study their properties is by building a *particle accelerator* — such as those at CERN in Europe (see Figures 20, 21, and 22) and Fermilab in the USA (see Figure 23).

Units for Measuring the Energy of Fundamental Particles

When atomic physicists learned how to isolate electrons, they found that the normal units of physics — such as the gram for mass and the joule for energy — were inconveniently large for discussing electron properties. They therefore introduced another unit — the *electron volt*, (symbol eV) — which was defined as the work required to move an electron through a potential difference of 1 volt. An electron volt is a small amount of energy — similar to the amount of energy posessed by a single photon of light.

Experiments in particle physics involve higher energies than those in atomic physics. Typical energies are of the order of 10^6 eV (a million eV, symbol MeV), 10^9 eV (a billion eV, or giga electron Volt, symbol GeV) or even 10^{12} eV (a trillion eV, or tera electron Volt, symbol TeV). Cosmic rays can possess even higher energies. The most energetic cosmic ray ever recorded had an energy of 320 EeV (an exa electron Volt, symbol EeV, is 10^{18} eV); this is the energy you would feel if you dropped a bowling ball onto your toe from a height of one meter.

Using Einstein's equivalence between mass and energy, $E = mc^2$, we can use eV/c^2, MeV/c^2, or GeV/c^2 as a unit of mass. Physicists do indeed use these units, but for convenience they drop the c^2, assuming the reader knows it is there. So, for example, you will often read that the mass of the proton is 938 MeV or that the mass of the Z boson is 92.6 GeV. I will follow that convention throughout this book.

In a particle accelerator, we start with electrically charged objects — usually protons or electrons — and use powerful electromagnetic fields to accelerate them to high energies and keep them focused in tight beams. In a linear accelerator such as SLAC (the Stanford University Linear Accelerator) the charged objects follow a straight-line path. In a circular accelerator such as the Large Hadron Collider at CERN and the Tevatron at Fermilab, a magnetic field is used to bend the charged objects in a circular or spiral path. We then smash the objects into a carefully prepared target, or crash them together in head-on collisions and examine the debris in special detectors (see Figure 24). A myriad of different objects are created in the collisions, and these can then be studied and their properties recorded.

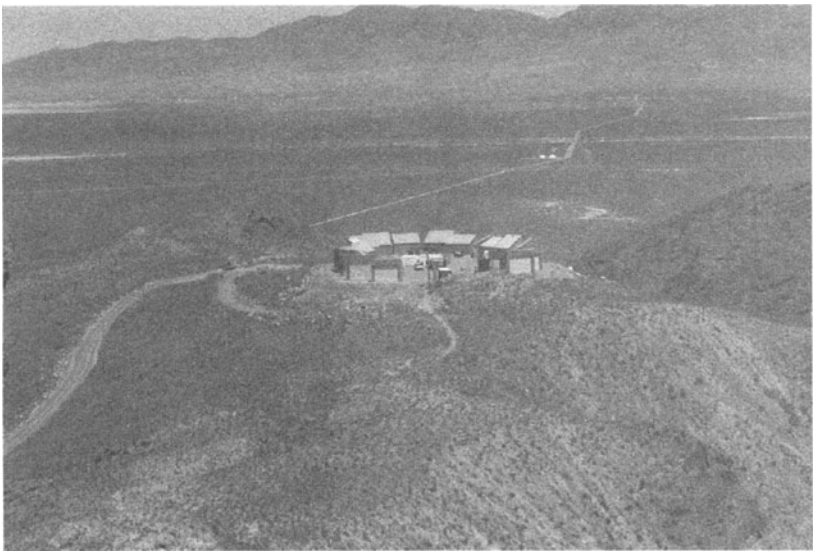


FIGURE 19 An aerial view of HiRes2, located at Dugway Proving Ground in Utah. It is an experiment that detects cosmic rays in a volume of about 30,000 cubic kilometers of atmosphere. The energy of the primary cosmic ray can be as high as 100 million TeV. The primary interacts in the atmosphere and produces an air shower of up to 10 billion charged particles. These secondary particles excite nitrogen molecules in the air, which then radiate ultraviolet light. Sensitive detectors then measure this ultraviolet light above the night sky background.

Through the careful analysis of the properties of these subatomic objects, physicists have discovered which are the truly fundamental particles and which are composite objects. For example, the proton was long thought to be fundamental; now we know it contains more basic entities.

All this may sound straightforward, but the practice is not so clean. A popular experiment involves colliding a proton (which, as we will see, consists of three quarks) against an antiproton (which consists of three antiquarks) in the hope of getting a quark to scatter off an antiquark. This is rather like firing two oranges together at hypersonic speeds and, in the midst of all the mess, hoping to see one seed bounce off another. Recording and then recognizing an interesting particle is a challenge that stretches technology to the limit. It seems like even more of an impossible challenge when you appreciate the tremendous precision of the experiments taking place at these colliders. Certain experiments at CERN have been so sensitive that physicists had to take note of the position of the Moon. (The varying gravitational pull of the Moon as it orbited the Earth had a measurable effect on the energy of the particle beam.) They even had to take note of the train timetable. (High-speed trains travelling out of Geneva bound for Paris had an effect on the beam energy.)

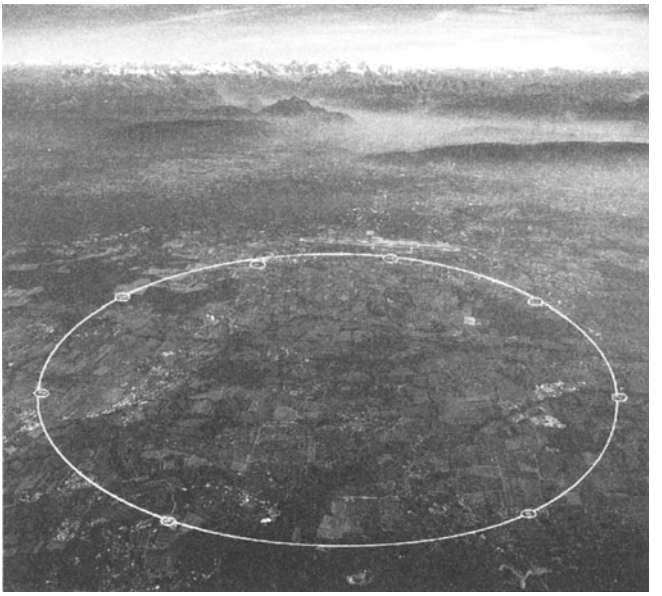


FIGURE 20 The Large Hadron Collider, due to become operational in 2007, uses an existing accelerator tunnel (shown here as a white circle superimposed on the landscape). The tunnel is 27 kilometers long and lies an average of 100 meters below Earth's surface. It is situated near Lake Geneva and the Jura mountains.

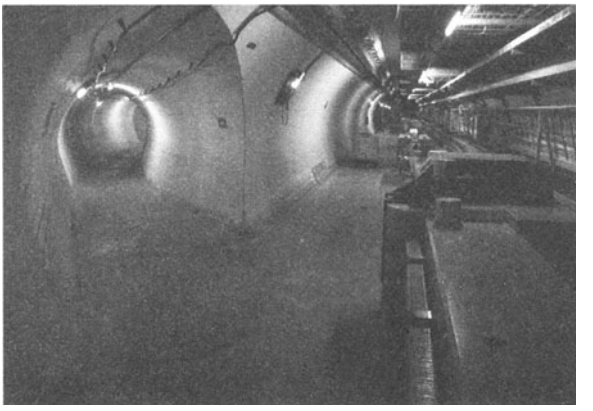


FIGURE 21 One of the CERN tunnels, deep underneath the Jura mountains.

Particle accelerators are huge structures, the size and complexity of which dwarf many industrial factories. This is "big science" at its biggest. Such large constructions are needed to accelerate subatomic particles to higher and higher energies. The greater the energy of collision, the more costly and difficult the experiment — but it is worth going to these energies for several reasons. The most important reason has to do with resolution. We use accelerators to probe short distances (in order to find out whether an object contains smaller objects and, if so, what those objects are). In other words, accelerators are essentially large microscopes. As a rough guide, however, a microscope can only resolve distances compara-

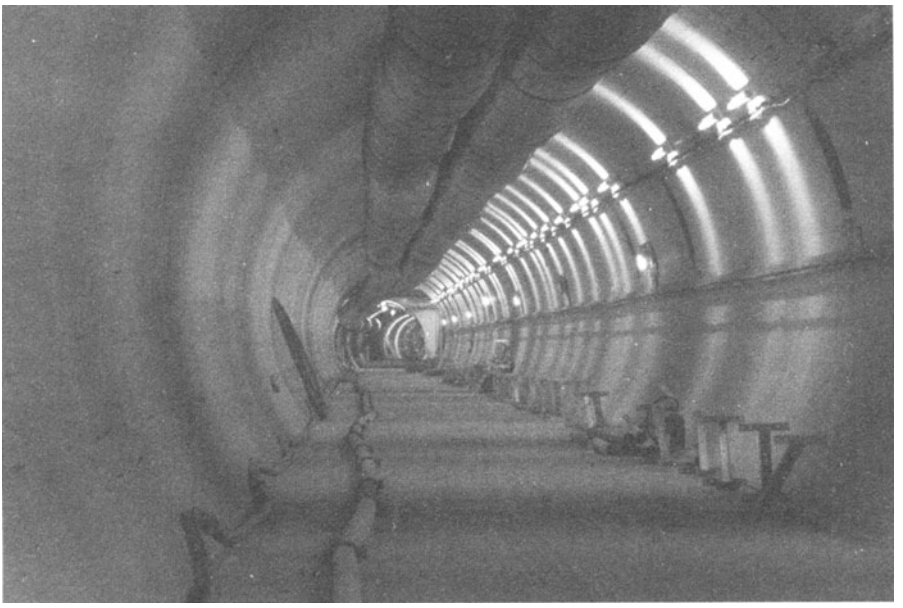


FIGURE 22 The Large Hadron Collider is pushing several technologies, including computing, superconductors, and electronics, to the limit. When it switches on, it will collide protons at an energy of 14 TeV, and it will crash lead nuclei together at a collision energy of 1150 TeV. Four giant detectors, called ALICE, ATLAS, CMS, and LHCb, will record the outcomes of trillions of particle collisions; during operation, ATLAS and CMS will each be recording data at a rate roughly equivalent to the contents of 10,000 copies of Encyclopaedia Britannica per second.

ble to the wavelength that is being used. In an optical microscope, the wavelength is that of light; in particle accelerators, the wavelength is the quantum wavelength associated with the particle. But quantum mechanics tells us that a particle's wavelength and its momentum are inversely related. So the more momentum a colliding beam can transfer, the shorter the associated quantum wavelength, and the finer the resolution we may achieve. That means we need to go to higher and higher energies.

The uncertainty principle tells us what resolution we can expect from our particle accelerators. For example, if colliding beams can achieve a momentum transfer of 10 GeV, then we can expect a spatial resolution of 10^{-16} meters: our "microscope" could examine distances that are smaller than the radius of the proton. (It was not until the early 1930s that accelerators capable of reaching 1 MeV were built, and not until the early 1950s that the 1 GeV barrier was broken. So in the early days of experimental particle physics, the resolution that could be achieved was low. At these low resolutions, the proton appeared as a point particle, which is one reason why the proton was incorrectly assumed to be a fundamental particle.)

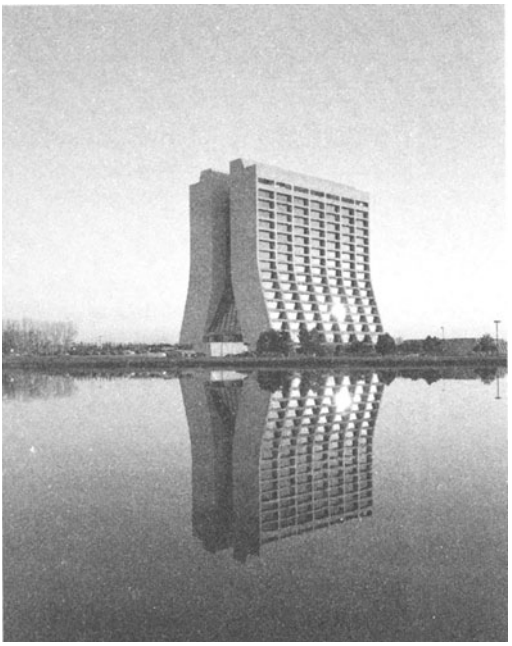


FIGURE 23 Wilson Hall, which was inspired by a Gothic cathedral in Beauvais, France, is here shown reflected in Swan Lake at sunset. This building, in Batavia, IL, is the focal point for administrative and scientific activity at Fermilab. The business end of Fermilab physics takes place at the Tevatron — an accelerator that is four miles in circumference and is, at the time of writing, the world's highest-energy particle accelerator. Three of the fundamental particles were first observed at Fermilab: the bottom quark, the top quark, and the tau neutrino.

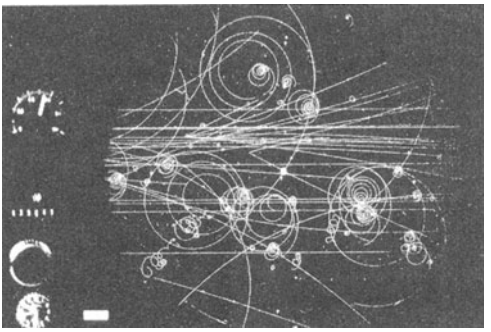


FIGURE 24 Once particles have been produced, we need to detect them. Until the 1970s, bubble chambers were often used to detect particles. A bubble chamber is a vessel filled with a fluid that is on the verge of boiling; when a charged particle passes through the vessel, the particle generates a string of bubbles — due to boiling — along its trajectory. The bubbles can then be photographed and analyzed. This is a CERN bubble-chamber photograph of the production of jets of particles. Each track in the photograph is the record of a short-lived charged particle, spiralling in an electromagnetic field. This photograph was taken in about 1970. Bubble chambers have long been replaced by more modern — though less aesthetically pleasing — methods of detecting particles.

Figure 26 shows the correspondence between energy and distance. Current high-energy experiments at CERN and Fermilab have probed distances down to 10^{-18} meters. As we shall see in the next chapter, our theories of fundamental particles work extremely well at this distance scale. Many of the "wild ideas" we shall discuss in this book, however, apply

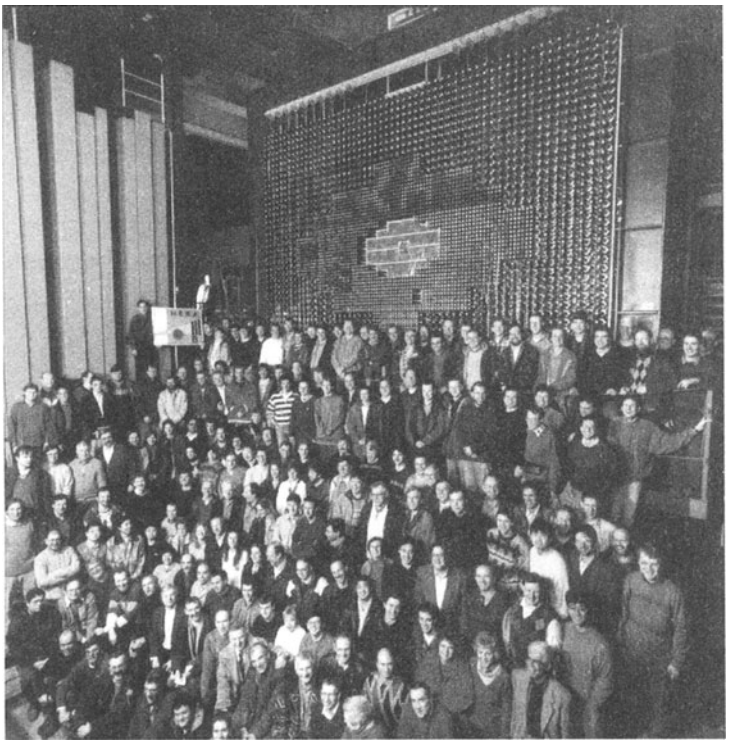


FIGURE 25 Particle physics experiments are nowadays so complex and expensive that they cannot be managed on a national level. Large international teams are required for the design, construction, and operation of the accelerators and detectors. This photograph shows the HERA-B team of scientists, who work on an experiment at the DESY laboratory in Hamburg. The experiment studies subatomic particles called B mesons. About 250 physicists, coming from 33 institutes in 13 countries, work on HERA-B. This is by no means the largest collaboration in particle physics.

only at ultra-high energies and correspondingly ultra-small distances — Planck-scale physics, in other words. It is unlikely that we will ever build an accelerator that can probe distances at the Planck length (10^{-35} meters). But that does not mean that the ideas are untestable. Astronomers have deduced that the Universe began in a singular event known as the Big Bang. The Universe started out extremely small and the energies of particles in the immediate aftermath of the Big Bang were incredibly large. The wild ideas we shall be discussing, if they have any validity, must have applied in the very early history of the Universe. So by studying cosmology we may be able to find traces of a time when the Universe was in a quite different state, and thus gain a window on these radical ideas. And, in Chapter 11, we will see how completely new physics may occur at energies that are indeed accessible to our particle accelerators.

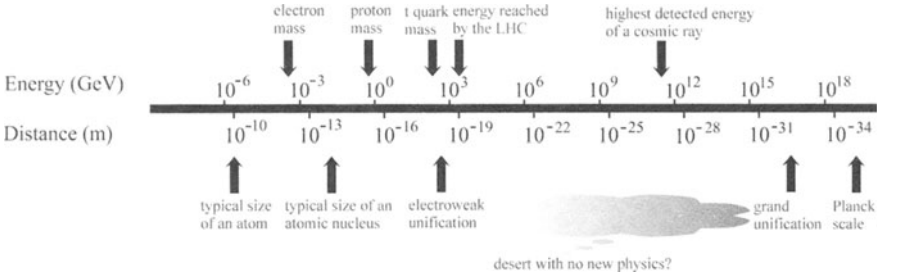

electron mass
proton mass
t quark mass
energy reached by the LHC
highest detected energy of a cosmic ray
Energy (GeV)
10^{-6} 10^{-3} 10^{0} 10^{3} 10^{6} 10^{9} 10^{12} 10^{15} 10^{18}
Distance (m)
10^{-10} 10^{-13} 10^{-16} 10^{-19} 10^{-22} 10^{-25} 10^{-28} 10^{-31} 10^{-34}
typical size of an atom
typical size of an atomic nucleus
electroweak unification
grand unification
Planck scale
desert with no new physics?

FIGURE 26 As we move to higher and higher energies, we probe nature at smaller and smaller distance scales. As we shall see in Chapter 4, current theories work extremely well in describing physics up to an energy of about 200 GeV (about 10^{-18} meters). At higher energies, or shorter distances, some new ideas must come into play. The received wisdom has been that no new physics will show up until we reach about 10^{17} GeV. Recent ideas, however, suggest that we will find interesting new physics at about 10^{3} GeV. This is the energy region that will soon be explored by the Large Hadron Collider at CERN. The ultimate goal is to describe physics at the Planck scale of 10^{19} GeV (about 10^{-35} meters).

The Web and the Grid

It is perhaps no surprise — to physicists anyway — that the World Wide Web was invented at CERN. By the late 1980s, a system was needed to provide a human-friendly interface to the vast amounts of information generated by research groups there. To this end, an English computer scientist named Tim Berners-Lee wrote the hypertext transfer protocol (http), the language computers use to communicate hypertext documents over the Internet, and designed a scheme to give documents addresses on the Internet. He also developed a browser to retrieve and view hypertext documents.

The next generation of particle collider at CERN — the Large Hadron Collider — will come on line in 2007, and the trillions of particle collisions that will take place there will generate more than 10^{7} Gigabytes of data each year (equivalent to the contents of about 20 million CD-ROMs). The analysis of this avalanche of data will require truly immense amounts of computing power. That is why CERN scientists are now working on the Grid. This is a new approach to computing that will work in much the same way as the electrical grid: we will plug into a computational grid to obtain computing power just as we now plug into the electrical grid to obtain electrical power. This will allow scientists to work with more computing power than any single machine can reasonably possess. Like the Web, the Grid will at first be available only to universities and research laboratories; but also like the Web, the Grid will soon be available to the rest of us. The new applications that the Grid will make possible will surely transform our world.

CLASSIFYING THE FUNDAMENTAL PARTICLES

To the small part of ignorance that we arrange and classify
we give the name of knowledge
Ambrose Bierce, *The Devil's Dictionary*

Over the years, trillions of events have taken place in particle colliders. What makes possible the study of such a myriad of events (apart from raw computing power) is our old friend symmetry. We can make sense of the outcome of particle collisions only in terms of symmetry: we need the conservation laws in order to explain the events that we observe in the laboratory. As we saw earler, the external spacetime symmetries possessed by quantum fields give rise to conservation laws involving energy, linear momentum, and angular momentum. Every particle collision ever recorded has obeyed these laws. If a particular process is forbidden by a conservation law, then that process is never observed to happen. Conversely, if a process is allowed by the conservation laws, then eventually (and perhaps at great expense) we *will* observe it!

Symmetry is important in another way. As we shall soon see, fields have "internal" symmetries, and these symmetries give rise to "charges." Charges are strictly conserved — not just globally, but *locally*, at each and every spacetime point. Electric charge is the most familiar type of charge, and in every particle collision the net electric charge before the event is the same as after the event. The same is true for other charges. Symmetry thus helps us find order in what would otherwise be an impenetrable maze.

Finally, symmetry is important in allowing us to classify the fundamental particles.

The first person to appreciate the power of symmetry for classifying the fundamental fields was Eugene Wigner. In 1939, Wigner explored what happens when quantum field equations are made to obey Lorentz invariance — in other words, by demanding that the equations obey special relativity. He showed that the simple demands of invariance under various transformations imply that quantum particles, which are the manifestation of the fields, possess certain properties — such as mass, charge, and spin — that are not subject to the uncertainty principle. These properties are thus the logical way of classifying particles and fields.

Rest mass An important part of the Wigner classification scheme is that quantum particles possess a definite *rest mass*, which can either be positive or zero. A positive rest mass is the mass that you measure when the particle is at rest relative to you. (If a quantum particle has a non-zero rest mass then you can always, at least in principle, catch up to it — so that it is at rest relative to you — and then measure its mass.) If a particle has a zero

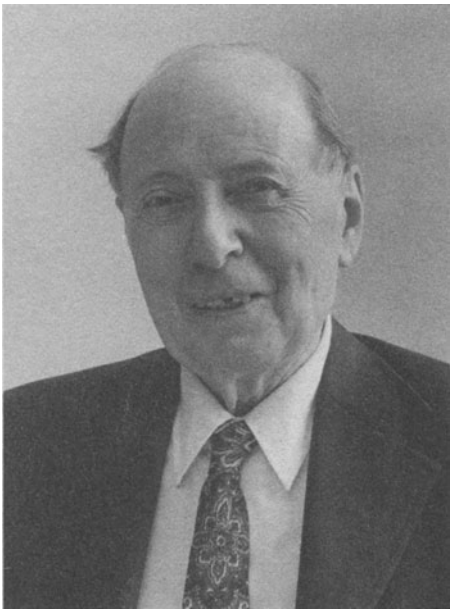


FIGURE 27 Eugene Wigner used the power of symmetry principles to help classify the fundamental fields. For this work, and his many other contributions to the study of quantum mechanics, he was awarded the 1963 Nobel prize for physics.

rest mass then it always moves at the speed of light, so you can never catch it. The photon — the particle manifestation of the electromagnetic field — has a zero rest mass and, of course, travels at the speed of light.

So quantum particles are either massless, or else they possess a definite rest mass.

Electric charge The Dirac field possesses a property called *phase*. You can think of the phase as an arrow that points from the origin to somewhere on a circle of radius 1. (See Figure 28.) We can pick *any* value for the zero point of the phase — any direction for the arrow — without affecting physics in any way. (This is rather like calculating the amount of energy it takes you to lift this book through a vertical distance of 12 centimeters. Within obvious limits, the energy you expend is independent of where you start. You can lift the book in your basement or on top of your roof and you will get the same result. The energy depends only on the difference in height, not on the choice of initial, or zero, height. You could say, therefore, that there is a symmetry under the choice of zero height.) Choosing the zero point for the field is called choosing a *gauge*. There is an invariance, a symmetry, under changes in the choice of phase for the Dirac field. It is an example of a *gauge symmetry*. This is an internal symmetry, which is quite unrelated to spacetime symmetries like the invariance of the field under rotations and translations. A continuous symmetry like this implies, via Noether's theorem, that something is conserved. And the conserved quan-

tity? Electric charge! The conservation of electric charge is a consequence of an abstract symmetry — invariance under choice of phase.

Electric charge thus gives us another way to classify particles. A particle either possesses an electric charge, or else it is neutral. Later, we shall meet various other types of charge, all of which arise because of particular symmetries of the fields. These charges also help us classify particles.

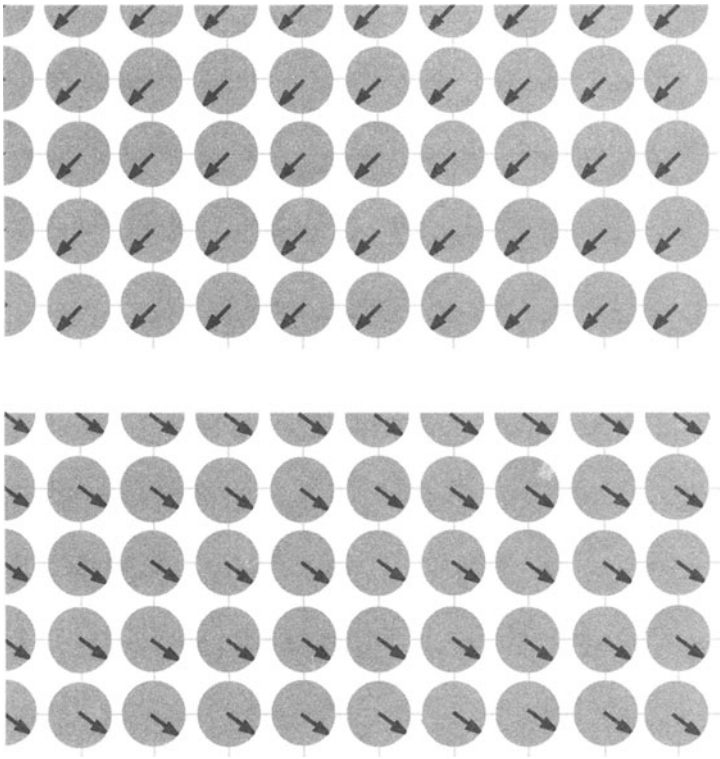


FIGURE 28 The Dirac field has a property called phase. Phase is represented here by an arrow pointing from the center of a circle of radius 1 and ending on the circumference of the circle. We are free to choose the direction of this arrow; in other words, we are free to choose a gauge. In the top part of the diagram, the arrows point to the lower left. In the bottom part of the diagram, the arrows point to the lower right. Nature is invariant under this transformation; there is no detectable difference between the two situations. We call this a global *symmetry, since the transformation has to be the same at all spacetime points. This invariance of the Dirac field under choice of phase gives rise, via Noether's theorem, to conservation of electric charge.*

Spin If we study how the fundamental quantum fields transform upon rotation, then we can assign them a definite *spin*. In units of $\hbar$ — which, as we saw on page 40, is the constant that governs quantum behavior — the Dirac field has spin ½; the associated quantum particle, the electron, thus

has spin ½. The electromagnetic field has spin 1; the associated quantum particle, the photon, thus has spin 1.

The spin of a particle is fixed forever. It does not slow down to a stop in the way that a flywheel eventually slows down, nor can you whip it into a faster spin state as you can with a spinning top. The spin of a particle is an inherent property of the particle. (We can think of a particle as spinning about an axis like a spinning top as long as we remember that the analogy is not perfect. The quantum world can never be perfectly pictured in terms of the properties of everyday objects.)

Wigner showed that quantum particles must possess either an integer spin (that is, spin 0, 1, 2...) or else a half-integer spin (that is, spin ½, 3/2, 5/2...). It is impossible for a particle to possess, say, spin ⅓ in quantum units (at least, it is impossible if special relativity is correct). Particles of integer spin are called *bosons* in honor of Satyendra Bose, an Indian physicist who collaborated with Einstein in investigating the subtleties of quantum mechanics. Particles of half-integer spin are called *fermions* in honor of Enrico Fermi, an Italian–American physicist and Nobel prize winner. Thus quantum particles possess another quantity that enables us to classify them. They are all either bosons or fermions.

Classifying particles as to whether they have integer or half-integer spin may seem trivial, but the different symmetry properties of these particles (see the boxed text on page 63) give rise to many important quantum phenomena. For example, an exclusion principle forbids identical fermions from occupying the same state. The existence of the periodic table — and ultimately the chemical properties of the world around us — depends upon fermions obeying the exclusion principle. If electrons did not obey the exclusion principle, then they would just pile into the lowest-energy state of an atom and matter as we know it would not exist. Bosons, on the other hand, obey no such principle. The quite different behavior of bosons is illustrated by phenomena such as laser beams. In old swashbuckling films, when pirates crossed swords it was perfectly reasonable to watch one sword stop another. Swords are material objects, which contain fermions that obey the exclusion principle, and hence will not penetrate each other. The fermions in the swords cannot be in the same place at the same time. In the film "Star Wars," when Luke crossed lightsabers with Darth Vader, the laws of physics were being broken. Lightsabers (presumably) are made of light beams, which consist of bosons that do not obey the exclusion principle, and hence cannot stop penetration. A beam of light will simply pass through another beam of light; lightsabers cannot work as shown in the film. Matter and radiation — fermions and bosons — behave quite differently, and our world looks the way it does because of that difference.

Fermions and Bosons

There is a profound difference between the behavior of bosons and fermions. This difference relates to symmetry.

Consider a pair of fundamental particles, one at position 1 and one at position 2. This two-particle system is described by some quantum wavefunction (the wavefunction, remember, tells us everything we can know about a quantum system). Suppose these particles have the same quantum numbers: the same spin pointing in the same direction, the same mass, the same charge; since these are fundamental particles with the same quantum numbers, they are *completely* identical. What happens if we swap the particle at 1 with the particle at 2? Well, we get a new wavefunction. The observable system, though, must be unchanged; the two particles are identical, so we cannot distinguish between the two situations. If we *could* observe a difference, then the particles would not be identical. This invariance under the transformation of swapping identical particles is known as exchange symmetry.

Now, in quantum mechanics, the probability of observing a system in a particular state is given by the square of the wavefunction. In the example above, the squares of the two wavefunctions must be identical. There are only two ways that the wavefunction can satisfy this constraint: either the wavefunctions are identical, or else one is the negative of the other. So the indistinguishability of fundamental particles means that wavefunctions are either *even* under the swapping of position (the wavefunctions are identical) or else they are *odd* under the swapping of position (one wavefunction is the negative of the other). It turns out that all fermions are odd under the exchange of identical particles. All bosons are even under the exchange of identical particles.

Suppose we try to put two identical fermions in the same place at the same time. The fermion wavefunction is odd under the exchange of identical particles, and this can be satisfied only if the wavefunction is zero. The probability of finding two identical fermions at the same place at the same time is zero. Identical fermions would not be seen dead together in the same place. This is the exclusion principle. When two identical fermions approach each other, a powerful repulsion comes into play. This is *not* an electrical repulsion: it happens between neutrinos, for example, which are electrically neutral. It is due, instead, to symmetry.

Bosons are different. We can put two bosons at the same place and the wavefunction need not vanish. Indeed, far from obeying a prohibition on like states occupying the same point in space, bosons *prefer* to be in the same state. This is the principle behind the laser, for example.

In addition to being able to measure the spin of a particle, we can also simultaneously measure the component of the spin along some particular direction (usually taken to be the z direction). For an electron, which is a spin-$\frac{1}{2}$ particle, the z-component of spin can either point in the positive direction of the z-axis (in which case it is said to be "spin up") or in the opposite direction (in which case it is said to be "spin down"). The direction of spin *is* something that can be changed; we can apply an electromagnetic field and flip an electron from a spin-up state to a spin-down state, and vice versa.

For a massive particle with spin 1, there are three possible orientations for the spin $(+1, 0, -1)$. The photon is a *massless* spin-1 particle, and it turns out that such a particle can possess only two orientations for the spin: $+1$ or -1. So photons can be said to possess a particular handedness — they can be left- or right-handed. (This effect is present in more familiar classical physics. In classical electromagnetism light is considered to be a transverse electromagnetic wave: electric and magnetic fields oscillate in a direction that is at right angles to each other and to the direction of propagation of the wave. Polarization describes the direction of the electric field, and there are just two separate states: vertical and horizontal. Common sources of light such as candles and torches emit unpolarized light, which means that the electric field vibrates in many different planes. But you can easily demonstrate that only two independent polarization states exist if you have have two polarizing filters — such as two pairs of Polaroid sunglasses. Hold one pair of glasses horizontally and the other pair vertically, directly in front of the first pair. No light gets through.)

A First Look at the Fundamental Particles

If I could remember the names of all these particles,
I would have been a botanist.
Enrico Fermi

According to our classification scheme, the truly fundamental particles are of two types: fermions and bosons. The role these two types of particle play will become clear in Chapter 4. For now, we can content ourself with a first look at particles that have been produced in colliders. A brief description should be enough to imbue the fundamental particles with some personality.

Let's start with a description of the fermions.

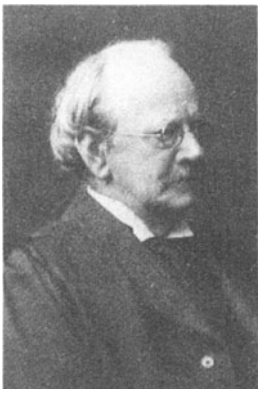


FIGURE 29 J.J. Thomson, Nobel Laureate 1906

The electron The electron — the first fundamental particle to be discovered in experiment — was found in 1897 by J. J. Thomson. In many ways, Thomson's discovery marks the birth of particle physics.

The electron's properties are well-documented. We know its rest mass (0.511 MeV), its electric charge (−1), its spin (½), and its other quantum properties. As with all the other apparently fundamental particles we shall be discussing, no experiment has detected any internal structure in the electron. It appears to be a point object. So once we have specified an electron's quantum numbers, we have said all there is to say about it.

Experimentally, electrons are easy to produce. It usually takes relatively little energy to strip an electron from its orbit around an atomic nucleus. Once an electron has been liberated, we can isolate it, study it, and manipulate it. The entire electronics industry, after all, depends on our ability to produce and then manipulate free electrons.

The up and down quarks It takes much more energy to liberate the particles in an atomic nucleus — the protons and neutrons — than it does to liberate an electron from the outer layers of an atom. This is because the force acting on particles in an atomic nucleus is much stronger than the electromagnetic force acting on electrons. During the 1940s, 1950s, and 1960s, physicists discovered dozens of subatomic objects that interacted via this strong nuclear force. Collectively these objects were called *hadrons*. (Particles such as the electron, which do not feel the strong force, were called *leptons*.) The hadrons fell into two types: *baryons* with half-integer spins, and *mesons* with integer spins. At first it was thought that baryons like the proton and mesons like the pion might be truly fundamental particles. But the role played by the menagerie of other hadrons — baryons like the lambda, sigma, and xi, and mesons like the rho, omega, and phi — was confusing.

The fog began to lift in 1964, when two American theorists, Murray Gell-Mann and George Zweig, independently introduced the quark model of the hadrons. The idea is straightforward. Baryons and mesons consist of quarks. It is the *quarks* that are the truly fundamental particles and, by appearing in various combinations, they produce all the known hadrons. In order to account for most of the hadrons that were known about in 1964, only two types (or *flavors*) of quark were required: the up quark (symbol u) and the down quark (symbol d).

Every baryon consists of a triplet of quarks (and every antibaryon consists of three antiquarks); every meson consists of a quark–antiquark pair. Simple. Beautiful.

Familiar hadrons contain only the u and d quark flavors. The proton, for example, consists of two u quarks and one d quark (uud). The quark content of the neutron is udd. The neutral pion contains one u quark and one u antiquark; the positively charged pion contains one u quark and one d antiquark; the negatively charged pion contains one d quark and one u antiquark. And so on. The same combination of quarks can also appear as "excited" states — states with more mass–energy. The Δ^+ baryon, for example, has the same quark content as the proton but is 30 percent heavier. The bewildering complexity of the particle tables produced by physicists in the 1950s was replaced in the 1960s by simplicity: all these objects were just allowed combinations of u and d quarks and antiquarks.

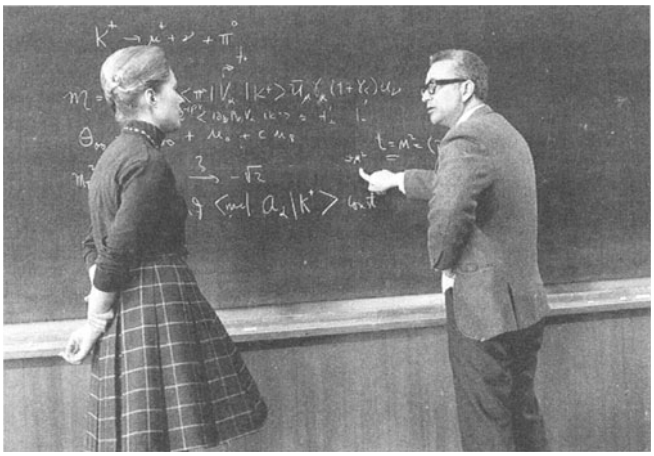


FIGURE 30 Murray Gell-Mann in conversation with Mary Gaillard. The list of Gell-Mann's contributions to physics is a long one — his development of the quark model of hadrons is just a small part. He is also an expert in linguistics and archaeology; some of his colleagues believe he has more in-depth knowledge of a wider range of subjects than anyone else alive. Gell-Mann was awarded the 1969 Nobel prize for physics.

The reason why quarks combine in only two ways — to form either baryons or mesons — will become apparent in Chapter 4. As we shall see there, the apparently unremarkable observation that baryons and mesons are the only allowed quark combinations leads to an important theory of nature. (Recently, nuclear physicists have found evidence for an object

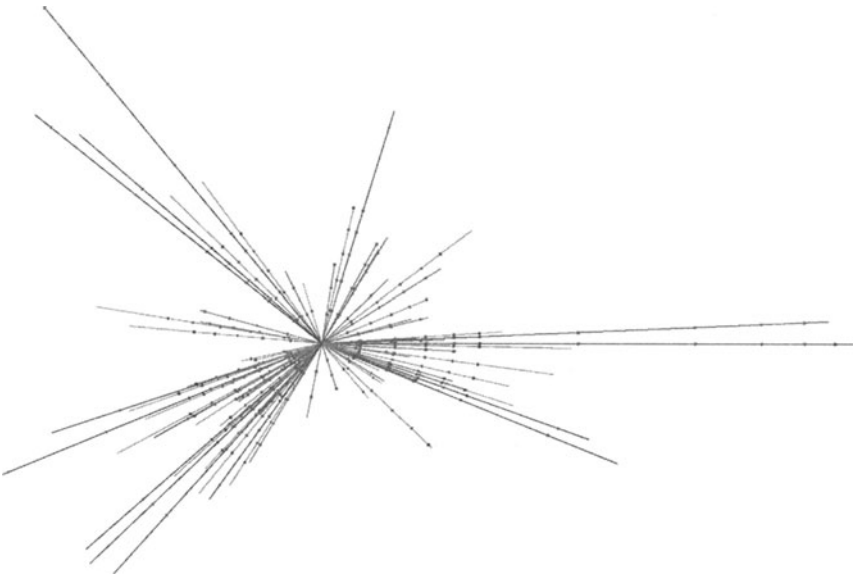


FIGURE 31 The ALEPH experiment at CERN studied collisions between electrons and antielectrons. When electrons and antielectrons come close enough they annihilate in a burst of energy that immediately rematerializes as matter in accord with Einstein's equation $E = mc^2$. In the collision shown in this diagram, the energy materialized as two quarks and two antiquarks. Quarks must "clothe" themselves; in doing so they form jets of particles. These jets shoot out into the detector from the interaction point, carrying the momentum of the original quarks. Although we may be unable to see isolated individual quarks, the jets of hadrons that we observe in many particle physics experiments are impossible to explain unless they are formed from energetic quarks.

containing four quarks and an antiquark. The so-called *pentaquark* is interesting, but we can perhaps think of it as a "molecule" of a baryon and a meson. Its existence does not cast doubt on the quark model of hadrons.)

It is important to appreciate that no one has ever seen a "free" quark, and there are good reasons to suppose we never will. Unlike electrons, which we can isolate with ease, quarks are confined within hadrons. Nevertheless, the evidence for quarks is compelling, and when we see jets of hadrons appear in experiments at particle accelerators (such as the jets shown in Figure 31) we are seeing the effects of individual quarks.

Since quarks do not appear in isolation, the notion of "quark mass" is slightly ambiguous. Under a commonly accepted definition of quark mass, the u quark has a mass of 4 MeV and the d quark has a mass of 7 MeV. (Recall that the proton mass is 938 MeV; clearly, most of the mass of the commonly observed subatomic objects is in the form of binding energy.) The u and d quarks are thus heavier than electrons — though only by a factor of about ten. An unusual feature of quarks is that they carry a

fractional electric charge. Taking the proton charge to be +1, the charge on the u quark is $\frac{2}{3}$ and the charge on the d quark is $-\frac{1}{3}$. (So the neutron, which is a udd triplet of quarks, has a charge $\frac{2}{3} - \frac{1}{3} - \frac{1}{3} = 0$. It turns out that *all* hadrons end up carrying either zero or integer electric charge; we never see a fractional charge, just as we never see a free quark.)

The electron neutrino The existence of the neutrino associated with the electron was postulated by Wolfgang Pauli in 1930 purely on the grounds of symmetry. An atomic nucleus can sometimes decay into a less massive nucleus by emitting an electron in a process called beta decay. The trouble was, energy seemed not to be conserved in such decays: the energy contained in the electron plus the decay nucleus was less than the energy of the original nucleus. In order to preserve the principle of conservation of energy, and thus the invariance of the laws of physics under translation in time, Pauli argued that another particle must be emitted along with the electron. This extra particle would carry the energy required by the conservation law. Such a particle would be insubstantial: it would have little or perhaps zero mass, and it would have zero electrical charge. It would, however, possess a spin: it would be a spin-$\frac{1}{2}$ fermion. It was Fermi who gave it the name *neutrino* ("little neutral one"); in this book we shall refer to it as the electron neutrino (symbol ν_e), since it is closely linked to the electron and we have to distinguish it from two other types of neutrino.

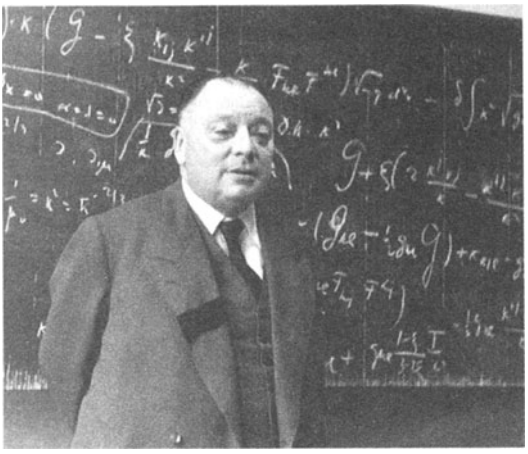


FIGURE 32 Wolfgang Pauli, Nobel Laureate 1945

Although the electron neutrino seemed to be necessary from a theoretical point of view, it was not until 1956 that Clyde Cowan and Frederick Reines actually detected the particle. The difficulty with neutrino detection from an experimenter's point of view is that the neutrino is unaffected by the forces that work on electrons, protons, and neutrons. It interacts so weakly with matter that it is difficult to stop a neutrino.

Modern neutrino detectors are typically located below Earth's surface in disused mines, or on seabeds; the shielding afforded by hundreds of meters of rock or water stops cosmic rays and other radiation from reaching the detectors — but neutrinos penetrate with ease.

Electron neutrinos were once thought to be massless. An important recent discovery is that they possess a tiny mass (less than 3 eV).

The electron and its neutrino form a natural pairing of fundamental particles. They are often represented together as (e, ν_e). The up and down quarks also form a natural pairing of fundamental particles. They are often represented together as (u, d). Together, these two pairings form a *family* of fundamental particles.

Most of the matter we see around us ultimately consists of this family of four particles: the u and d quarks combine to form protons and neutrons, which in turn combine to form nuclei. Nuclei combine with electrons to form electrically neutral atoms, which combine to form elements and chemical compounds. The presence of electron neutrinos is far less obvious, but they are formed in vast numbers in nuclear reactions deep inside the Sun (billions of solar neutrinos are streaming through you as you read this sentence). It seems that most of the matter in the Universe requires representatives from only one family of fundamental particles. Yet for some reason, this family is replicated twice over.

The muon and the muon neutrino The muon (symbol μ) was discovered in 1936 in cosmic-ray experiments performed by the American physicist Carl Anderson and his graduate student, Seth Neddermeyer. (Four years earlier, Anderson had discovered the antielectron — the antimatter partner of the electron — the existence of which had been predicted by Paul Dirac. Anderson won the Nobel prize this work.) The muon neutrino (symbol ν_μ) was first identified experimentally in 1962 by a trio of American physicists — Leon Lederman, Melvin Schwartz, and Jack Steinberger — in an experiment at Brookhaven. The three physicists shared the 1988 Nobel prize for the discovery.

Decades of study have determined that the muon is 207 times more massive than the electron. (Its rest mass is 105.659 MeV.) Otherwise, it seems to be identical to the electron: it has the same spin, the same electric charge and, like the electron, it has an associated neutrino. (The mass of the muon neutrino is not known, but it does not exceed 0.19 MeV.)

So in addition to the (e, ν_e) pairing of leptons, we have the (μ, ν_μ) pairing. No one knows *why* nature has ordered this extra set of particles. But they exist, and they are as fundamental as the more familiar electron and electron neutrino.

The strange and charm quarks Imagine if Cleopatra had fallen from her barge in 50 BC and had not yet hit the Nile. That was the conundrum facing physicists in the 1950s when they came across subatomic particles with an unusually long lifetime. In 1953, Murray Gell-Mann, and independently Kazuhiko Nishijima and Tadao Nakano, introduced the concept of "strangeness" to solve the conundrum. (See the boxed text below.) This was more than a decade before the quark idea itself was formulated, but the concept is readily accommodated in the quark model.

Strangeness

Fifty years ago, the timescales on which phenomena were observed to happen was used to distinguish between the various types of interaction. Strong interactions between hadrons typically occurred on timescales as short as 10^{-23} seconds. Weak interactions typically took place on longer timescales — up to 10^{-8} seconds. The timescale for electromagnetic processes was somewhere in between.

Experiments had shown that certain hadrons could be produced in collisions on a very short timescale — about 10^{-23} seconds. These same hadrons, however, stayed around for a long time before decaying — as long as 10^{-8} seconds. (One hundredth of a microsecond may not seem like a long time but, compared to the timescale on which these hadrons were produced, it was an eternity.) So these hadrons were produced by the strong interaction but decayed through the weak interaction. This discovery was a problem for physicists. They could think of no reason why a hadron that was formed by the strong interaction should not also decay by the strong interaction. The behavior of these hadrons was anomalous, or strange.

The explanation for their behavior was that these hadrons possessed a property — called strangeness — that was conserved in strong interactions but not in weak interactions. For example, the collision of a pion and a proton, each with a strangeness of 0, could produce via the strong interaction *two* strange particles with opposite strangeness (and thus a total strangeness still of 0). However, suppose the produced particles were the lightest type of strange particle. Then their decay could not be by the strong interaction, since there would be no less massive strange particles into which they could decay. The decays would instead have to happen via the much slower weak interaction, which does not conserve strangeness.

Strangeness, S, is a property possessed by certain hadrons. So in addition to the u and d quarks, there must be another flavor of quark, the

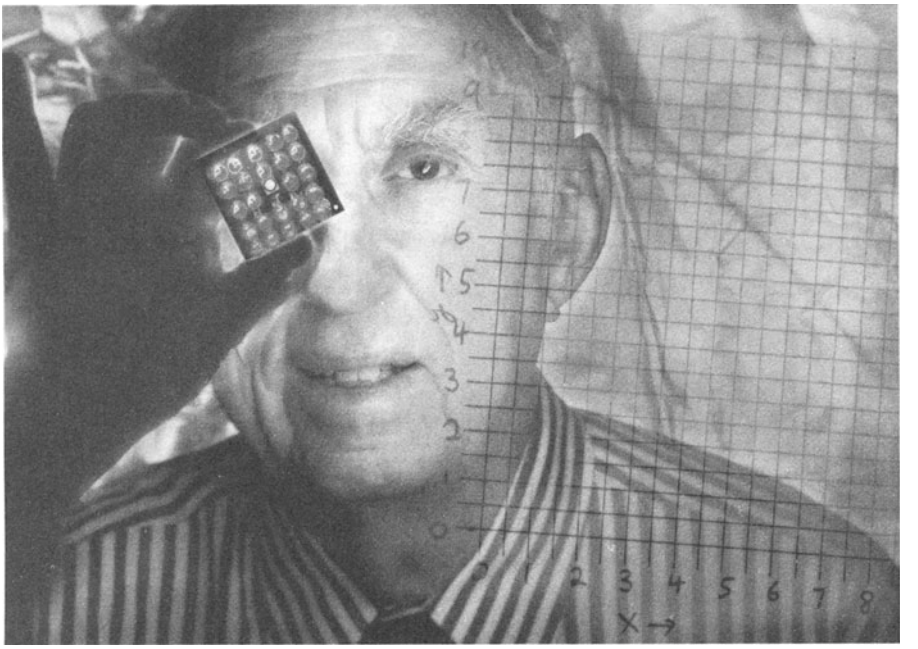


FIGURE 33 Martin Perl shared the 1995 Nobel prize for physics with Frederick Reines. Perl won the prize for his discovery in 1975 of the tau lepton. His discovery signalled the existence of a third family of particles; the bottom quark was found in 1976, the top quark in 1995, and the tau neutrino in 2000. It now seems highly likely that there is not a fourth family of particles; all quarks and leptons have been discovered.

s quark, which carries a strangeness of $S = -1$. The corresponding anti-quark of course carries $S = +1$. The u and d quarks carry a strangeness of 0. According to the quark model of the hadrons, if a baryon contains one s quark, then it has a strangeness of -1 (an example of such a baryon is an object called the Σ^-, which has a quark content of dds). If the baryon contains two s quarks, then it has a strangeness of -2 (an example is the Ξ^-, with a quark content of dss). With three s quarks, it has a strangeness of -3 (this is the Ω^- baryon, with a quark content of sss). Clearly, $S = -3$ is the maximum strangeness allowed for a baryon in the quark model. In all the particle collisions ever observed, no one has ever spotted a baryon with a strangeness greater than three; if the quark model is correct, no one ever will observe such a particle.

The mass of the s quark is about 100 MeV, and it carries an electric charge of $-\frac{1}{3}$.

Similar observations of other long-lived particles led physicists to believe that certain hadrons possess yet another quality, this time called charm.

This is accommodated within the quark model by assuming that there is a fourth flavor of quark, called charm (symbol c). Experimental evidence for the c quark came in 1974. The mass of the c quark is about 1200 MeV, and it carries an electric charge of $\frac{2}{3}$.

The two pairings of leptons — (e, ν_e) and (μ, ν_μ) — are thus matched by two pairings of quark — (u, d) and (c, s). We have a second family of fundamental particles.

The tau and tau neutrino Much of what was written above for the muon applies to the tau (symbol τ), which was discovered in 1975 by a team led by Martin Perl working with data from the Stanford Linear Accelerator Center. Perl's discovery came as rather a surprise, since it clearly signalled the existence of a third family of particles — something that theorists had been speculating about, but not very seriously. The rest mass of the tau was later measured to be 1784 MeV, which is about 3500 times more massive than the electron.

With the discovery of the tau, the search began for its associated neutrino. The existence of the tau neutrino (symbol ν_τ) was confirmed in 2000 by an international collaboration of 54 physicists working on an experiment called DONUT (Direct Observation of the Nu Tau) at Fermilab's Tevatron. The tau neutrino is more massive than the other neutrinos (its exact mass is unknown, but arguments suggest it must be less than 18.2 MeV).

We thus have *three* pairings of lepton: (e, ν_e), (μ, ν_μ) and (τ, ν_τ).

The bottom and top quarks Much of what was written above for the s and c quarks applies similarly to a third pairing of quarks: the top (symbol t) and bottom (symbol b) quarks. The (t, b) quark pairing thus matches the (τ, ν_τ) lepton pairing and we have a third family of fundamental particles.

The b quark was discovered experimentally in 1977 by a team of physicists led by Leon Lederman at Fermilab. Lederman's team discovered a meson called the upsilon, which consists of a b quark and its antiquark. The b quark has a mass of about 4.2 GeV and, like the d and s quarks, carries an electric charge of $-\frac{1}{3}$.

The t quark, which was discovered in 1995 in two different experimental setups at Fermilab's Tevatron, weighs in at about 180 GeV. This is similar to the mass of a gold nucleus, which contains 197 protons and neutrons. The t quark is the most massive of any of the fundamental particles yet observed. Like the u and c quarks, it carries an electric charge of $\frac{2}{3}$.

That concludes our brief survey of the fundamental fermions; now let's take a first look at the fundamental bosons. As we shall see in Chapter 4, the fundamental bosons play a different role than the fundamental fermions. The bosons act as carrier particles of the fundamental forces.

The photon The photon (symbol γ) is the carrier particle of the electromagnetic interaction. It acts on any particle that possesses electric charge.

The photon is certainly the most familiar of all the fundamental particles. We see light; we feel radiant heat; we cook with microwaves; we have our bodies examined with X-rays. All of these phenomena involve photons. We may give photons many different names — radio waves, visible light, gamma rays, and so on — but the name we give them simply depends on their frequency (and thus their energy). Deep down, all these seemingly different phenomena are the same. They all involve the emission and absorbtion of photons.

The photon is remarkably common in the Universe: for every triplet of quarks that form a proton, there are about 10 billion photons. Explaining this ratio of photons to protons is an important task for cosmologists and particle physicists.

The W and Z bosons We have seen that the quarks come in six different types, or flavors: up, down, strange, charm, bottom, and top. The charged leptons come in three different types: electron, muon, and tau, each of which has a corresponding neutrino. Sometimes, in radioactive decays or in particle collisions, flavor can change. For example, a down quark can transform into an up quark and a muon can transform into an electron (in both transformations, other particles are involved too). The fundamental force that is responsible for these flavor-changing processes is called the weak force, and the carrier particle of that force comes in two types, the W boson and the Z boson.

The weak force affects all the fundamental fermions, both leptons and quarks, and is responsible for some of the most important processes in the cosmos. If there were no the weak interactions, the chain of nuclear reactions that generates energy in the Sun would break. The Sun would not shine and life as we know it would probably not exist.

The W boson is electrically charged, and in the usual units it carries one unit of charge (the charge can be positive or negative, so there are W^+ and W^- bosons). The W boson is extremely heavy — it weighs in at 81.8 GeV.

In addition to the flavor-changing weak processes, there are some weak processes that do not change flavor. These processes are mediated by the electrically neutral Z^0 boson. The Z^0 boson is even more massive than the W boson — it has a mass of 92.6 GeV. In many ways, though, the Z boson acts as if it were simply a more massive version of the photon. In fact, as we shall see in Chapter 4, when we discuss the various interactions in more detail, there are very close links between the photon and the W and Z bosons.

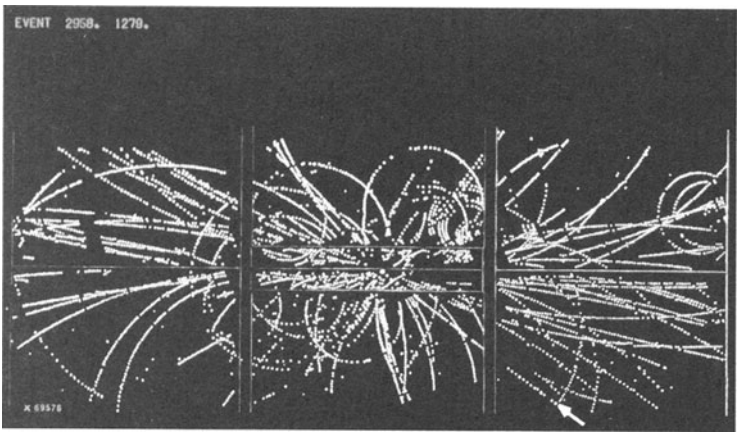


FIGURE 34 One of the events found by the UA1 detector in its October–December 1982 run. The event produced an electron (indicated by an arrow in the lower-right part of the photograph) with a high transverse energy. The electron was produced back-to-back with "missing energy" — the explanation for which was that an antineutrino was emitted. The only interpretation of this event that made sense was that a W boson had been created and then decayed into an electron and an electron antineutrino.

The Discovery of the W

Although theorists had predicted the existence of the W boson as far back as the 1960s, the difficulty in producing particles of such high mass (the W boson is almost 90 times more massive than the proton) meant that the W boson was not detected until 1982 (see Figure 34).

FIGURE 35 In 1976, the Italian physicist Carlo Rubbia (pictured here) had the idea of converting an existing accelerator into a storage ring for protons and antiprotons. He hoped that collisions between the stored particles would have sufficient energy to detect the W and Z bosons. The Dutch physicist Simon Van der Meer invented a method for storing protons and antiprotons that made Rubbia's idea feasible. The W and Z particles were found, and both men were awarded the 1984 Nobel prize for physics.

Since their discovery, several properties of the W and Z bosons have been measured to an accuracy of a few parts per thousand. There is good agreement between these measurements and predictions from theory.

Z bosons are produced by colliding beams of electrons and positrons at just the right energy. Z bosons decay to produce either a quark and its matching flavor antiquark or a lepton and its matching antilepton. These decays are readily detectable.

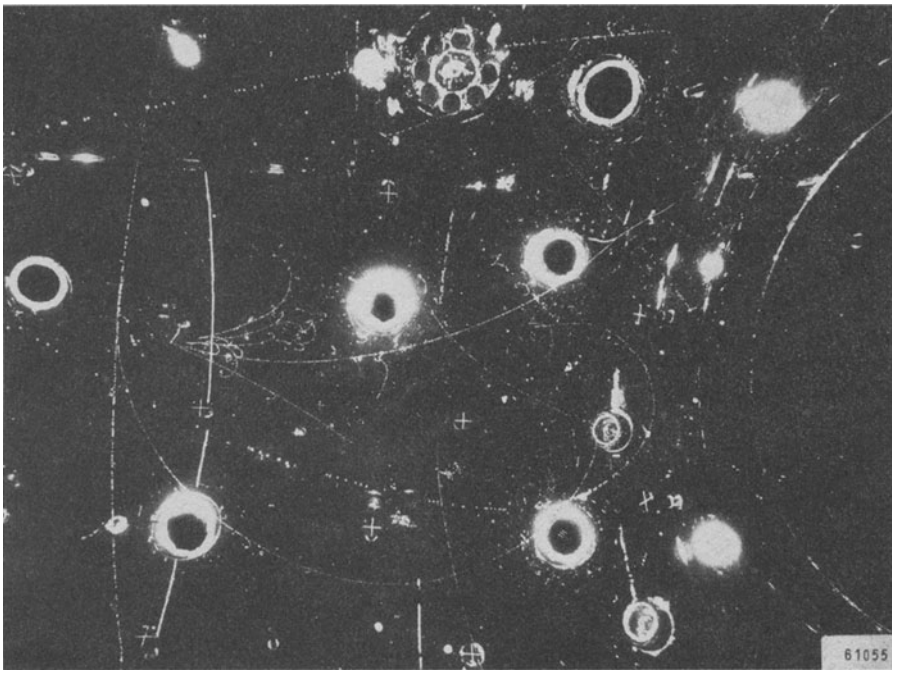


FIGURE 36 This photograph, first published in 1973, was taken in a CERN bubble chamber called Gargamelle. A beam of muon neutrinos passed through the freon-filled chamber; the neutrinos themselves left no tracks. However, a spray of particles appeared "from nowhere," emerging from a point towards the center-left of the photograph. These particles were the debris of a freon nucleus that was struck by one of the energetic — but invisible — neutrinos. Apart from the gravitational interaction that all particles feel, neutrinos interact solely via the weak interaction. In a weak interaction such as this, we would normally expect the muon neutrino to convert into a muon; but no muon appeared in the debris — the neutrino continued on its elusive way, unchanged. The only explanation was that the neutrino interacted via the exchange of a neutral particle — the Z^0. Photographs like this were the first experimental proof that neutral weak interactions were possible and were important because the existence of the Z^0 was a clear prediction of theories that unified the weak and electromagnetic interactions. (It was another ten years, however, before physicists were able to produce the Z^0 particle directly.)

The gluon In the 1950s, physicists thought that the strong force was something that acted between protons and neutrons and bound them together to form atomic nuclei. We now know that the nuclear forces acting on protons and neutrons is just a remnant of a much stronger, more fundamental force — the force that "glues" quarks together to form the protons

and neutrons themselves. The carrier particle of this strong interquark force is called the gluon.

Gluons, like quarks, cannot be produced in isolation. But, as with quarks, the evidence for the existence of gluons is overwhelming. In 1979, the detection of "jets" of particles following the collisions of electrons and positrons led physicists to conclude that they had observed the clear effects of gluon production.

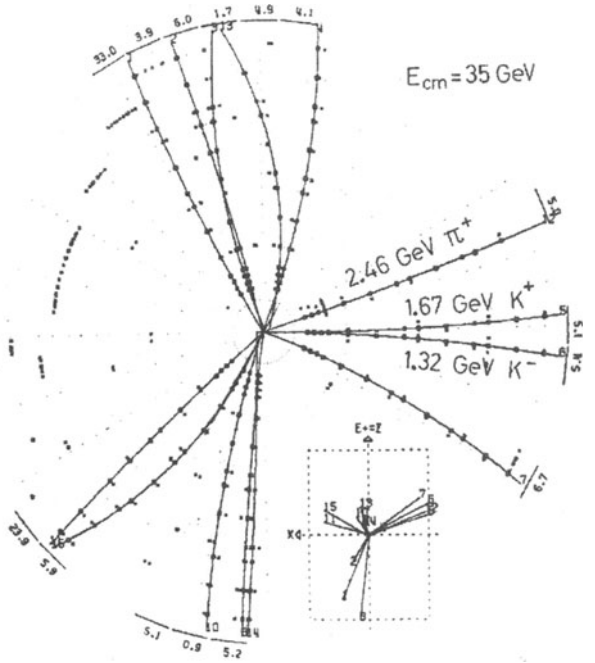


FIGURE 37 Gluons are the carrier particle of the strong force. They were observed for the first time in 1979, at PETRA (an electron–positron collider at DESY in Hamburg). Gluons, like quarks, can never be isolated; they are confined within hadrons. Nevertheless, the observation of "jets" of particles with certain properties, as in this photograph, can only be explained in terms of gluons.

The gluon is in some ways rather like the photon. The photon "couples" to a charge — the electric charge — and carries the electromagnetic force. The gluon also couples to a charge — the color charge, as we shall describe in the next chapter — and carries the strong force. However, a major difference between the gluon and the photon is this: the photon is electrically neutral, but the gluon possesses a color charge. In fact, there are eight gluons, each of which carries a different color charge. The gluon, because it is charged, can couple to itself. This is quite unlike the behavior of photons. It leads to the possibility, for example, of *glueballs* — completely new states of matter that consist of pure glue.

This concludes our tour of the basic building blocks of the Universe. As Figure 38 shows, the number of fundamental particles is rather small. It is astonishing that the interactions of such a few basic entities can generate the beauty and intricacy of the world we see around us.

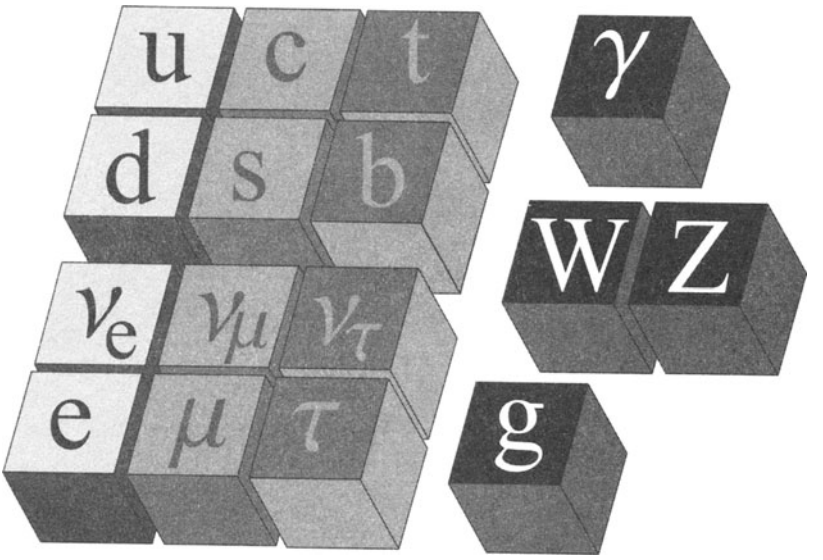


FIGURE 38 The building blocks of the Universe. The top two rows of blocks represent the six quarks (up, down, charm, strange, top, and bottom). The bottom two rows of blocks represent the six leptons (electron neutrino, electron, muon neutrino, muon, tau neutrino, and tau). Each of the quarks and leptons has a corresponding antiparticle. Note that the quarks and leptons are also arranged in columns; each column represents a family of particles. The blocks to the right represent the force particles: the photon, the weak bosons W and Z, and the gluon. From these basic building blocks we can construct a model of particle physics.

4

THE STANDARD MODEL

There is certainly no absolute standard of beauty.
That precisely is what makes its pursuit so interesting.
J. K. Galbraith

Our present understanding of the microworld is defined in terms of quantum fields and the interactions between them. The best description we have of the fields and their interactions is given by the Standard Model of particle physics.

The Standard Model is one of the crowning achievements of twentieth century science. It is the pure distillation of more than one hundred years of investigation by physicists, and the theoretical framework is based on the firmest of experimental evidence. Indeed, at the time of this writing, the Standard Model can explain the results of *every* experiment that has ever probed the behavior of matter on the smallest scales. Only gravity remains

outside the remit of the Standard Model. (In particle physics experiments we are dealing with very tiny masses and the gravitational effects are completely negligible.) This chapter introduces some of the key concepts of the Standard Model. By understanding these ideas, and appreciating their tremendous success we are then in a better position to understand why the Standard Model must be incomplete. We will better understand what drives physicists to contemplate wild ideas such as large extra dimensions and colliding universes.

Our Best Theory

Never worry about theory as long as the machinery does what it's supposed to do.
Robert A. Heinlein

The Standard Model is a theory — or rather a set of theories — that encapsulates and explains all the experimental studies of the microworld that have ever been made. All the phenomena of Maxwell's electromagnetism, all the studies of radioactivity, all the data produced in particle accelerator laboratories — all of this can be explained by the Standard Model. The Standard Model is the end result — so far — of what happens when we keep asking "why" questions about the fundamental material of the Universe and how that material interacts. Remarkably, for such a wide-ranging framework, the Standard Model is extremely simple; it contains just two basic components.

First, there are some fundamental particles — the quanta of the fundamental fields. The particles are the "words" of the Standard Model. To recap, there are two types of fundamental field: fermions and bosons. The fermions are of two types: three families of lepton (the electron and its neutrino, the muon and its neutrino, and the tau and its neutrino) and three families of quark (the up and down quarks, the charm and strange quarks, and the top and bottom quarks). Each of the 12 fermions has an antiparticle. The fundamental bosons, all of which possess spin 1, are the photon, the three weak bosons (W^+, W^-, and Z^0), and the eight gluons. The bosons serve as their own antiparticles. From these few basic building blocks, all the familiar matter in the Universe can be constructed.

Second, these fundamental particles interact in only three ways. (Remember that we are ignoring gravity.) The interactions are the "grammar" of the Standard Model. Fundamental particles can interact via electromagnetism which, as we shall see later, is an aspect of the unified electroweak interaction. Electromagnetism acts on any particle that carries electric charge — every particle except the neutrinos, in other words; the car-

rier particle is the photon. The fundamental particles can also interact via the weak interaction, another aspect of the unified electroweak interaction. The weak interaction acts on all particles (it is the only interaction felt by the neutrinos) but its effects are often masked by the other interactions; the carrier particles here are the W and Z bosons. The fundamental particles can also interact via the strong interaction. The strong interaction acts on any particle that carries color charge. (In other words, it acts only on the quarks and gluons.) It is the force that binds quarks to form the hadrons; the carrier particle in this case is the gluon. Each of the carrier particles couples to the charges with a characteristic "stickiness" known as the *coupling constant* for the interaction. The coupling constants are different, so the strengths of the three interactions are different.

In the previous chapter we looked at some of the experimental considerations that have driven the development of the Standard Model — namely, the production and detection of the fundamental particles. That gave us a vocabulary. In order to better understand the structure of the Standard Model, though, we need to take a look at the interactions; we need to understand some of the grammar.

In the everyday world, we can identify many different types of force — friction in automobile engines, drag and lift on aircraft, tensile forces in buildings, and so on. But at the quantum level of quarks and leptons only three forces — or interactions — are truly fundamental. Other forces are but dilute remnants of the electromagnetic, weak, and strong interactions. The nuclear force, for example, binds protons and neutrons in the atomic nucleus; but this force is merely what is leftover from the strong interactions between quarks in the various nuclear particles. On a larger scale, when a bat hits a ball the impact is ultimately governed by the electromagnetic interactions of electrons in the bat and the ball.

At the quantum level, the description of interactions is different from our everyday description of forces. Consider the interaction between two electrons. Electrons are constantly spitting out virtual photons — a phenomenon that is allowed by the uncertainty principle. The virtual photon emitted by one of the electrons may be absorbed by the other electron, transferring both energy and momentum between the electrons in the process. Using classical physics we might say that the electrons repelled each other, but at the quantum level the force is mediated by the exchange of a virtual photon. Ultimately, the electromagnetic interaction takes place via the exchange of virtual photons; similarly, the weak interaction is mediated via the exchange of virtual W and Z bosons, and the strong interaction is mediated via the exchange of virtual gluons. (In a sense, then, the fermions are "particles of matter" while the bosons are "particles of force.")

Hints of Physics Beyond the Standard Model

The Standard Model is in agreement with all experiments that have investigated the microworld. Recently, however, experiments have hinted at effects that may lie beyond the Standard Model.

A promising approach for probing the Standard Model is to identify predictions that experimenters can test with high precision. If experiment and theory do not match, it may be that the Standard Model is incomplete. At present, perhaps the best example of this approach involves muons.

If an electron (or muon) moves in a uniform magnetic field that is perpendicular to the direction of travel, the electron (or muon) will follow a circular orbit. Now, electrons and muons have an intrinsic spin. Suppose the spin axis initially points in the direction of motion; what happens to it as the particle completes a circular orbit? Simple quantum mechanics predicts the spin axis will continue to point in the particle's direction of motion. In the 1940s, however, experiments showed that the spin direction of an electron changes slightly as a magnetic field guides it around a circle. The explanation for this anomalous behavior required the development of quantum electrodynamics, or QED (see page 84). As an electron moves, it continuously emits and reabsorbs short-lived virtual particles and this affects how its spin changes direction; QED tells us by exactly how much. For an electron moving in a magnetic field, the agreement between observation and theory is staggering: they agree to ten decimal places.

A muon also changes its spin direction in a magnetic field due to emission and reabsorption of virtual particles. But muons are more massive than electrons so they are more likely to emit a massive particle in one of their quantum fluctuations. The effect of emitting and reabsorbing a Z boson, for example, is about 40,000 times greater on the muon than the electron. This fact may open up a window on physics beyond the Standard Model. We can calculate what happens to the muon's spin direction as it circles in a magnetic field, assuming only Standard Model particles interact with it. If we compare the results of this calculation with observation and find they are incompatible, it may be because massive fundamental particles exist that interact with muons but are not part of the Standard Model. (These may be supersymmetric particles, which we discuss in Chapter 5.)

Recently this muon experiment has been performed at the Brookhaven National Laboratory in Upton, New York. The measurements are larger than theory predicts. However, although the difference between theory and experiment is statistically significant it is also rather small. The experiment must run for several more years before physicists can say conclusively they are observing effects that lie beyond the Standard Model.

In practice, it is impossible to understand the impact of a bat on a ball in terms of the interactions of trillions of electrons. It is no easy matter even to understand the nuclear force in terms of the interactions of a few dozen quarks. But at particle accelerators we can study the simplest case: the interaction of one particle with another via the exchange of mediating force particles. It is through such experiments that physicists have been able to understand the interactions, and thus to develop the Standard Model.

The behavior of the fundamental interactions is described completely by two relativistic quantum field theories. *Quantum chromodynamics* describes the strong interaction, while the *electroweak theory* describes the electromagnetic and weak interactions in a unified manner. (Thus, at a deep level, there are only *two* interactions that are relevant for fundamental particles: the electroweak and strong interactions. To begin, though, we will look at the three interactions separately.) Not only do the interactions all take place via the exchange of virtual bosons, but the very existence of the bosons is demanded by the requirements of symmetry. There is thus a beautiful and common structure to the theories. So before we look at the interactions, let's look at the role that symmetry plays here.

Local Gauge Symmetry

Measure not the work,
Until the day's out and the labour done,
Then bring your gauges.
Elizabeth Barrett Browning, *Aurora Leigh*

In the previous chapter we considered phase as an example of gauge symmetry. The symmetry transformation described there was a *global* symmetry: pick a particular zero-point of phase *here*, and it must have the same zero-point *there*, and over all of spacetime (see Figure 28). Suppose we allow different spacetime points to pick a different zero point of the phase, as in Figure 39. In other words, suppose that each observer at a spacetime point can pick a local gauge. This seems reasonable; after all, why should a choice of gauge on a distant planet affect the observations we make here? But surely, in that case, all hopes of a finding a symmetry break down? Wondrously, it turns out that the symmetry is restored if we introduce another field — called a *gauge field* — to compensate for the rate of change of the phase with position or time. With perfect symmetry, the gauge field is massless and thus there is a massless gauge boson associated with it. For the Dirac field, we can identify what this new gauge field must be: it is the

electromagnetic field — and the massless gauge boson is the photon! In other words, by demanding that the Dirac field obeys *local* gauge symmetry, we are forced to accept the existence of the electromagnetic field. This is a wonderful illustration of the power of symmetry principles in physics. By demanding that various fields obey a local gauge symmetry, and following the logic of that demand, we eventually end up with a complete quantum theory of all phenomena involving electric charges; the theory is called *quantum electrodynamics* (QED). Successive gauge transformations of the Dirac field form a group, the gauge group $U(1)_{EM}$. (The subscript is just to remind us that this U(1) group, which could be a symmetry group of many different kinds of objects, has been applied to the problem of electromagnetism.) So in the lingo of the trade, we say that QED is a $U(1)_{EM}$ gauge theory.

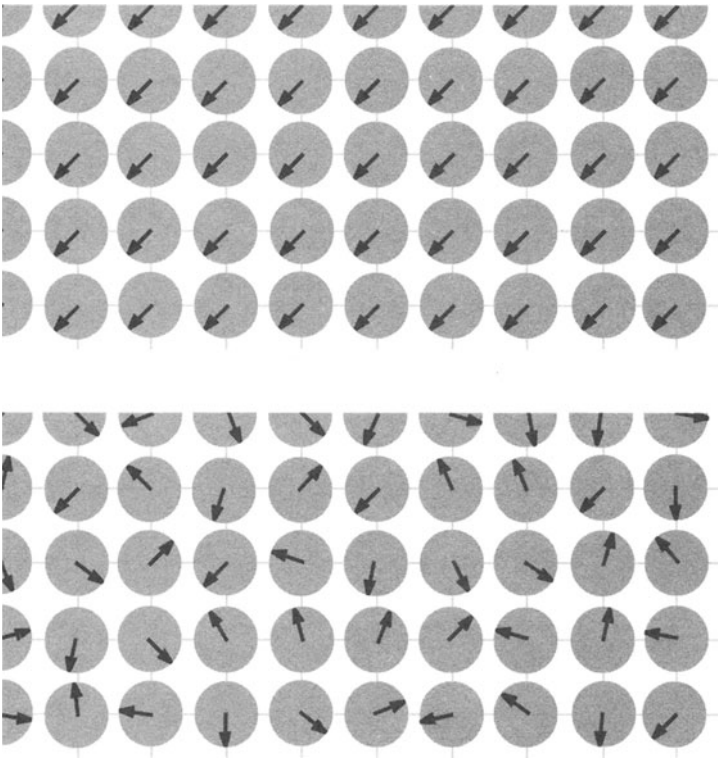


FIGURE 39 We again represent the phase of the Dirac field by an arrow pointing from the center of a circle of radius 1 and ending on the circumference of the circle. We are free to choose the arrow direction, and in the top diagram all observers have chosen the same direction. But observers at different points are free to make a local *choice of phase, as in the lower diagram. Remarkably, after this latter transformation, physics is unaltered provided another field — a gauge field — compensates for the rate of change of phase with position or time. This is a local gauge symmetry. For the Dirac field, the gauge field is the photon.*

As we shall see, the other interactions possess a local gauge invariance. We are free to choose a gauge at each point in spacetime, as long as a compensating gauge field carries information about the choice that has been made. Again, it turns out that successive gauge transformations form a group — the SU(2) group for the weak interaction and the SU(3) group for the strong interaction. So the Standard Model is a set of relativistic quantum field theories based on the idea of local gauge symmetry. Since QED is the simplest such theory and is the theory upon which the others are modelled, it makes sense to look at QED in a little more detail.

THE ELECTROMAGNETIC INTERACTION

God runs electromagnetics by wave theory on Monday, Wednesday, and Friday, and the Devil runs them by quantum theory on Tuesday, Thursday, and Saturday.
Sir William Bragg

Quantum electrodynamics was developed in the 1930s and 1940s, but calculating even a simple quantity — such as the probability of one electron scattering off another — was difficult to perform in the early formulations of the theory. It was Richard Feynman, a wonderfully imaginative physicist, who simplified the methods of doing calculations in QED. To aid him with the mathematics, Feynman developed a diagrammatic tool, now called the *Feynman diagram*. A Feynman diagram represents a particular physical process. For example, a basic process in electromagnetism is the scattering of two electrons off each other, which we represent as follows:

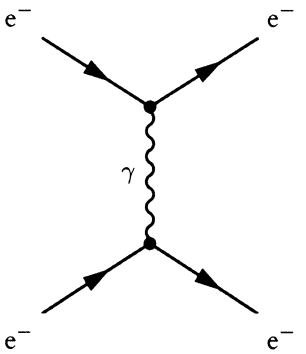


FIGURE 40 A Feynman diagram representing the scattering of one electron (symbol e^-) off another electron via the exchange of a photon (symbol γ).

We can interpret Figure 40 quite easily, as long as we remember two things. First, time increases as we read across the diagram. Second, the diagram is symbolic. The lines do *not* correspond to actual particle trajectories, such as the tracks left by particles in a bubble chamber (so they are

not like the tracks in Figure 24). Bearing these points in mind, the diagram says two electrons enter, they exchange a photon, and two electrons leave. In the process, the electrons transfer energy and momentum — they feel a force. The virtual photon is not observable; we see only the incoming and outgoing electrons. In classical physics, if the two electrons were initially at rest, we would describe this process by saying that "like charges repel." Here, we say that like charges interact via the exchange of a photon.

Feynman diagrams are useful not just because they help us visualize quantum processes, but because they represent a number. In essence, a Feynman diagram represents the probability of a particular quantum event happening. Feynman developed a set of rules for assigning a particular factor to each element of a diagram — the incoming and outgoing lines have certain factors applied to them; the internal photon has a particular factor applied to it; the vertices, where the fermions emit or absorb a photon, have factors applied to them; and so on. Once the factors have been correctly assigned within a diagram, we can then calculate the number associated with that diagram. You simply have to follow the recipe that Feynman wrote down. Now follows the important bit.

Suppose you want to analyze a particular physical process — the scattering of two electrons off each other, for example. Feynman said that you should draw all possible diagrams with the relevant external lines for that process (in this case two electron lines enter and two electron lines leave), calculate the number associated with each diagram, and then add up all the numbers. This result tells you the probability of the process occurring. You can then go away and compare the outcome of the calculation with actual experiments.

There are two problems with this approach, one of which is obvious, the other much less so.

FIGURE 41 Richard Feynman was a hugely influential figure in the physics community. Several books about his exploits — including Surely You're Joking, Mr Feynman! *and* What Do You Care What Other People Think? *— also made him a household name. He shared the Nobel prize for physics in 1965.*

The first problem is that we are told to draw all possible diagrams that contribute to a process. However, for any particular process there are an infinite number of Feynman diagrams!

For example, in the case of two electrons scattering off each other we can imagine the electrons exchanging a virtual photon, which then converts into an electron–antielectron pair before annihilating back into a photon. Such shenanigans happen constantly at the quantum level. Or perhaps the scattering electrons exchange *three* photons. Or perhaps the photon produces an electron–antielectron pair which annihilates back into a photon which produces an electron–antielectron pair before annihilating back into a photon. In each case, the same *observable* process is happening — two electrons enter, two electrons leave — but the internal mechanism of the interaction is quite different. So how can we calculate the probability of the process happening when there is an infinite number of contributions to the process?

Fortunately, one of the rules that Feynman devised says that we must assign to each vertex a factor that is proportional to the square root of α. The number α (which has the value $1/137$) is the coupling constant for electromagnetism — the famous *fine structure constant* — and is a measure of the strength of the electromagnetic force. Since the fine structure constant is less than one, the more vertices that a diagram possesses, the smaller will be its contribution to the final result. Figure 40 is an example of a so-called tree diagram: it contains only two vertices. The diagram that represents the process of two electrons exchanging a virtual photon, which then converts into an electron–antielectron pair before annihilating back into a photon, contains an extra pair of vertices. It thus has another factor of $1/137$ attached to it. It will be roughly 0.7 percent as large as the tree diagram. So we do not *have* to calculate every single higher-order diagram in order to get an accurate result. The first few orders will do. Quantum electrodynamics is an example of a *perturbative* theory: the higher orders that quantum mechanics says must exist are merely small perturbations on the main lowest-order result. (Of course, if we want to increase the accuracy of a calculation, and if we have the requisite stamina and patience, then we are free to calculate the contribution from higher-order diagrams.)

This is such an important concept, both here and in later chapters, that it is worth stating it again in slightly different language. We have in quantum electrodynamics a field theory that we can write down on paper in a few lines of mathematics. Unfortunately, the theory is too difficult for us to solve analytically. We do not know how to obtain nice mathematical expressions for the behavior of electrons in even simple situations. The best tool we have in such a situation is perturbation theory, in which we calculate successively better approximations to the answer. Nature has been

kind to us with QED: since the coupling constant is small, a perturbative approach works extremely well.

The second problem with QED is that there are diagrams like this:

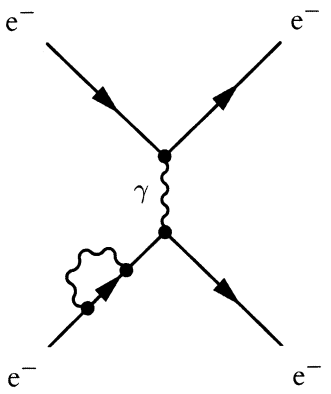


FIGURE 42 Here, one of the electrons emits and then reabsorbs a photon. When we calculate the number associated with this Feynman diagram we get an infinite answer. This infinity arises as an infinite contribution to the electron mass.

The problem is that when we calculate the number associated with the diagram in Figure 42, we obtain a value that is infinite. At each order of perturbation there are several diagrams that give an infinite contribution. Some of the greatest physicists of the twentieth century struggled for years with the infinities plaguing quantum electrodynamics. The solution, devised by Feynman and independently by Julian Schwinger and San-itiro Tomonaga, is called *renormalization*. The idea is that these infinities arise because of very small distance contributions of the electron structure. The infinities tell us that at very short distances (or, which is the same thing, at ultra-high energies) the theory breaks down and must be replaced by other ideas. This in itself is not a problem; theoretically we *expect* that QED will eventually break down, and practically we cannot probe these ultra-high energies. The difficulties arise because the infinities complicate the calculation of quantities that *are* accessible to us. What Feynman, Schwinger, and Tomonaga realized was that these troublesome QED diagrams only ever affect the electron mass, the electron charge, and the fine structure constant. But we know the values of these quantities from experiment. So we can sweep up the infinite contributions into definitions of quantities called the "bare" mass, "bare" charge, and "bare" coupling constant. The bare mass, charge, and coupling constant are infinite; but that need not concern us since we only ever work with a *physical* mass, charge and coupling. In essence, if we use the *physical* values in the Feynman rules, then we can ignore all the diagrams that give infinite contributions.

There are only a few theories in which *all* the infinities arising from higher-order diagrams can be dealt with in this manner; in other words, there are only a few theories that are renormalizable. Fortunately, gauge

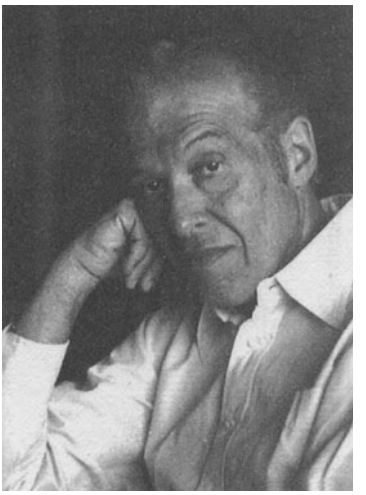


FIGURE 43 Julian Schwinger, along with San-itiro Tomonaga and Richard Feynman, shared the 1965 Nobel prize for physics for the development of quantum electrodynamics.

theories are renormalizable. Quantum electrodynamics, being a gauge theory, is thus a renormalizable field theory.

Many of the ideas of QED apply to QCD, the quantum field theory of the strong interaction. So let's look at that next.

THE STRONG INTERACTION

More strong than all poetic thought.
Alfred Tennyson, *In Memorium*

Quantum chromodynamics describes the dynamics of color charged particles just as QED describes the dynamics of electrically charged particles.

The basic idea is that quarks (but not leptons) carry a color charge, which is analogous to electric charge. The color charge, however, comes in three types: red, green, and blue. The quarks are not *really* colored, at least not in the everyday sense of the word. They possess a quality — a charge — that we could just as well call "snarkness" or "boojumness" or whatever. But the charge has certain properties that behave in the way that color acts in our everyday world, so we can live with the name.

So the fundamental quark fields possess a local gauge symmetry. In QED, nature is unchanged if we make an arbitrary choice of phase at each spacetime point as long as we introduce a massless gauge field that transmits the information about the choice made. In QCD, nature is unchanged if we make an arbitrary choice of color at each spacetime point as long as

Color

Physicists first introduced the idea of color for reasons that had nothing to do with the strong interaction. The initial idea merely provided a way out of a dilemma in Gell-Mann's quark model of the hadrons. According to the original version of the quark model, some of the baryons consist of more than one quark in the same state. The Δ^{++}, for example, is a uuu baryon with all three quarks in the same state. But the exclusion principle tells us that it is impossible to have more than one identical fermion in the same state. So what is going on? Well, if the quarks possess a new property — color — then the three u quarks in the Δ^{++} can be in different states: one quark could be red, one quark green, and the other blue. They would no longer be identical particles, and the exclusion principle would not apply.

The Δ^{++} has a total color charge of one unit of red, one of green, and one of blue. If these were colors in the everyday world, then we would say that the overall color of the Δ^{++} is white. In other words, Δ^{++} is colorless. It was not long before physicists realized that nature might be obeying a principle: *all observable particles are colorless*. The principle would provide a simple reason for the quark combinations that were showing up in accelerator experiments. We can make a colorless combination of three quarks (equal amounts of red, green, and blue), we can make a colorless combination of three antiquarks (equal amounts of antired, antigreen, and antiblue), and we can make a colorless combination of a quark–antiquark pair (red and antired, green and antigreen, and blue and antiblue). In other words, we can see baryons, mesons, and their antiparticles — the observed hadrons. But we cannot make a colorless combination of, say, two antiquarks or a quark with three antiquarks — and sure enough, such quark combinations have never been observed.

we introduce gauge fields that transmit the information of color choice from one point to another. The symmetry that these transformations obey is described by the group $SU(3)_C$. The 3 arises because there are three different color charges and the subscript reminds us that this is a symmetry under the interchange of color. The quantum particle associated with these gauge fields is called the *gluon*. Thus a mediator, analogous to the photon in QED, is responsible for the strong force between quarks.

Quantum chromodynamics rapidly became the best theory that physicists possess for understanding hadron interactions. As with quantum electrodynamics, calculations in QCD are aided by using Feynman diagrams. For example, the lowest-order Feynman diagram representing the interac-

tion between quarks (which is responsible, among other things, for binding them together into hadrons) is:

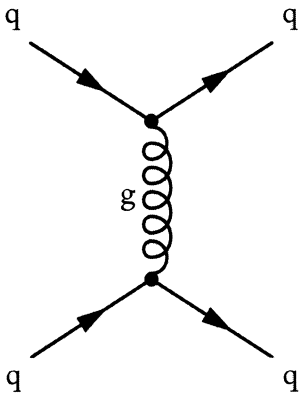


FIGURE 44 A basic quark–quark interaction.

Figure 44 looks just like the QED case of electron–electron scattering. Here, two quarks enter, they exchange a gluon, and two quarks leave. So presumably we simply calculate the contribution of the various diagrams involved in binding quarks and then we understand everything there is to know about hadrons. We will also understand the remnants of this force, which holds protons and neutrons together in a nucleus — terrific! Unfortunately, the situation is not so simple. There are three crucial differences between processes in QED and those in QCD.

First, there is only one type of electric charge. The charge can be positive or negative, of course, but only a single number is required to specify it. This means there is only one type of mediator for the electromagnetic interaction — the photon. Mathematically, this is a consequence of the $U(1)_{EM}$ symmetry of the theory. The color charge, though, comes in three varieties. The theory is described by the $SU(3)_C$ symmetry, and the mathematics of this group means there are *eight* types of gluon, all of which themselves carry the color charge. The gluons are an interesting mix of a color and an anticolor, and when a quark emits or absorbs a gluon, the quark can change color (though the quark flavor stays the same). For example, a red d quark might change into a blue d quark upon the emission of a red–antiblue gluon.

Quantum mechanics tells us that a quark does not have a *single* color that it keeps forever. On short timescales, quarks are ceaselessly emitting and absorbing gluons and so are constantly changing color. Figure 45 depicts how the colors of the quarks in a proton are constantly changing.

Gluons carry the color charge, so they are themselves subject to the strong force. In other words, gluons exchange gluons with other gluons,

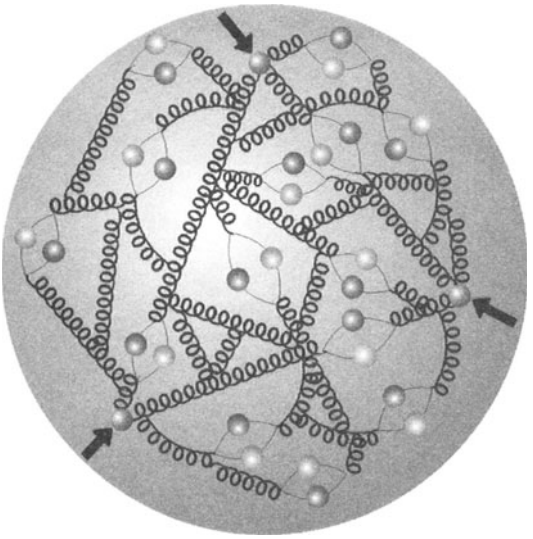


FIGURE 45 A proton contains three "valence" quarks (indicated here by arrows). These valence quarks are bound together by the exchange of gluons; the emission of a gluon can alter the quark's color but the proton overall remains colorless. The exchanged gluons can, on a short timescale, convert into quark–antiquark pairs, which can then recombine into gluons. A proton, then, is not just a collection of three quarks. It contains a bubbling froth of short-lived particles.

which allows for the possibility of colorless bound states of pure glue (called *glueballs*) and for colorless bound states of a gluon, quark, and antiquark (called *hybrid mesons*). It is clear that QCD is a much richer theory than QED, but it is correspondingly more difficult.

Gluon Colors

There are six gluons that can convert the color of a quark. Red–antigreen turns a red quark into a green quark; red–antiblue turns a red quark blue; green–antired turns a green quark red; green–antiblue turns a green quark blue; blue–antired turns a blue quark red; blue–antigreen turns a blue quark green. In addition, there are two different gluons that couple to the color charge on a quark without changing the quark color.

Second, the value of the strong coupling constant α_s — as deduced from experiments in nuclear physics — is greater than one. This is a major restriction for the theory. Remember that, according to the Feynman rules, a number proportional to the square root of the coupling constant appears at each vertex. In QED, the coupling constant is small, so each order of diagram contributes *less* than the previous order; we rapidly converge upon an answer. In QCD, each order of diagram contributes *more* than the previous order, and the answer diverges. The perturbation approach does not work.

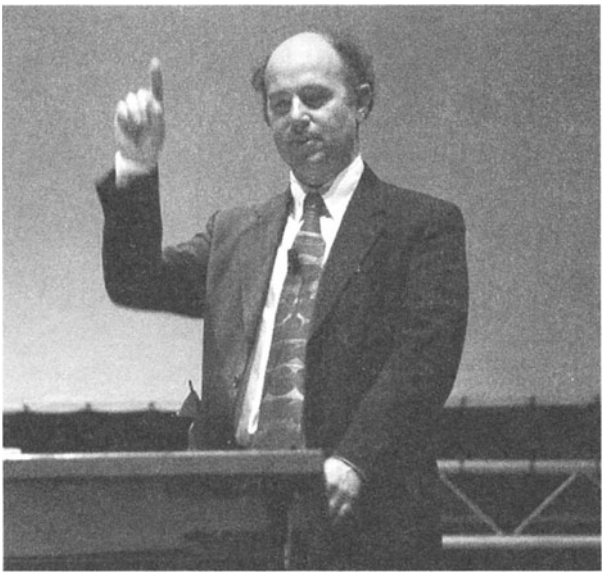


FIGURE 46 Frank Wilczek is a professor of physics at MIT. Along with David Gross and David Politzer he discovered that QCD possesses the crucial property of asymptotic freedom. This means that the force between two particles is strong when they are far apart, weak when they are close together.

Fortunately, there is a way around this problem — at least in some circumstances.

One consequence of renormalization is that coupling constants depend upon the energy at which we measure them. Physicists thus talk of "running" coupling constants. When we say that the QED fine structure constant α is $1/137$, for example, this is actually the value it possesses in the low-energy regime. If we performed experiments at higher energies, then we would measure a slightly larger constant. The value of the strong coupling constant also depends upon the energy at which we measure it; it turns out that the strong coupling constant runs much more quickly than the fine structure constant and in the opposite direction. At energies typical of nuclear physics, α_s is indeed greater than one and the perturbative approach fails. But at higher energies, when we start to probe distances much less than the size of the proton, α_s is smaller. Within a hadron, the force between quarks can be really rather small; the quarks rattle around as if they were free, and we are at liberty to use the perturbation approach. This property of QCD is known as *asymptotic freedom*, and it enables us to use the perturbation approach in some interesting areas of hadronic physics. (It was mainly the discovery of the asymptotic freedom of the theory that led to the general acceptance of QCD as *the* theory of the strong interaction. The discovery was made by David Gross and Frank Wilczek, then at Princeton University, and independently by David Politzer, then at Harvard University, who were working within a framework developed by the Dutch physicist Gerardus 't Hooft.)

Only at large separations does the force between quarks become large, and perturbation theory fails. It is as if the quarks were tied together by elastic bands: if the quarks are close together so that the bands are loose, there is little force between them. If the quarks are separated, the bands become stretched and the inter-quark force increases. This helps to explain why we do not see isolated quarks. Imagine a high-energy proton–proton scattering event, in which a quark in one of the protons strikes a quark in the other proton and knocks it away. The "elastic band" attaching this quark to the other two quarks will eventually break; but the result is *not* an isolated quark. There is so much energy contained in the band that quark–antiquark pairs or quark triplets can be created. The result is that the proton can maintain its status as a proton, but the escaping quark appears as a meson or a baryon. Sometimes there is so much energy in the "elastic band" that a whole shower of hadrons is created. The shower has the same momentum as the initial quark, so the hadrons appear as a jet of particles. Seeing these jets — as in Figure 31 — is almost as good as seeing an individual quark.

Although this elastic band explanation of quark confinement sounds plausible and is supported by numerical simulations done on computers, it is not *proof* that quarks are confined. Unfortunately, quark confinement occurs precisely in the regime where the coupling constant is large and the perturbative approach fails. To prove that quarks are confined we need to understand the non-perturbative aspects of QCD. At present this is beyond physicists — but as we shall see in later chapters, there are tantalizing hints that we are beginning to glimpse the non-perturbative theory. The hints involve higher dimensions and other "wild ideas" — but more of this later.

Third, a QCD process such as that illustrated in Figure 44 can never take place in the absence of other quarks. Only colorless states are observable in nature, so the quarks shown in Figure 44 would have to be accompanied either by an antiquark (to make a meson) or two other quarks (to make a baryon). We can only learn about QCD interactions by smashing *bundles* of quarks and antiquarks together in an accelerator; we can never do the simple, clean experiment of colliding a single isolated quark with another isolated quark. This is yet another reason why analyses in QCD are rather complicated.

In summary, QCD is a renormalizable relativistic quantum field theory based on the $SU(3)_C$ gauge symmetry, and in certain regimes we can apply perturbation theory to perform calculations. The technical problems involving QCD calculations mean that the agreement with experiment is much less impressive than is the case for QED, but there is every reason to suppose that QCD is the theory that describes hadronic physics. The fog surrounding physicists in the 1950s and 1960s, when new particles were

discovered on a monthly basis, has lifted. The underlying theory is simple, even if the details themselves are difficult.

THE WEAK INTERACTION

Made weak by time and fate.
Alfred Tennyson, *Ulysses*

Weak interactions act on all particles, both leptons and quarks. Furthermore, the weak interaction is the only one that can change the flavor of the particles involved in the interaction. Despite the ubiquity and importance of the weak interaction, though, weak processes are relatively rare. We observe them taking place in a system only if strong or electromagnetic processes are forbidden for some reason — for instance, if a conservation law prevents a strong or electromagnetic reaction from happening. Given the relative rarity of weak interactions, and the subtle and in some ways counterintuitive behavior they display, it is not surprising that it took physicists several decades to develop a successful quantum field theory of the interaction. We now know there are three mediating particles for the weak interaction: the W^+ and W^- carry an electric charge (the magnitude of which is equal to the electron charge) and the Z^0 is electrically neutral.

As before, we can draw Feynman diagrams to help us calculate quantities of interest. A typical weak diagram involving just leptons would be:

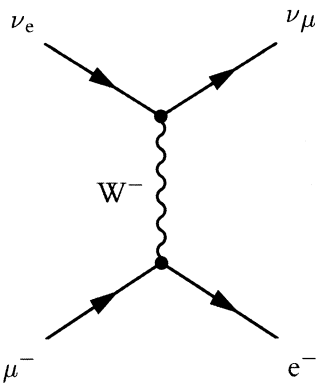


FIGURE 47 Muon–electron neutrino scattering. A muon (symbol μ^-) and electron neutrino (symbol ν_e) enter, exchange a W boson, and an electron (symbol e^-) and muon neutrino (symbol ν_μ) leave. By twisting this diagram, one obtains the more realistic process of a muon decaying into an electron, a muon neutrino, and an electron antineutrino.

The W boson couples to quarks as well as leptons. It does not couple to color — that is the job of the gluon — so the color of a quark is unchanged in a weak interaction. But the quark *flavor* can change in a weak process. The flavor is not carried away by the W particles, in the way that gluons carry away color; that cannot be the case, because the W particles couple

to leptons, and leptons do not possess flavor. So we must deduce that quark flavor need not be conserved in weak interactions. A typical example of a weak interaction involving both quarks and leptons would be:

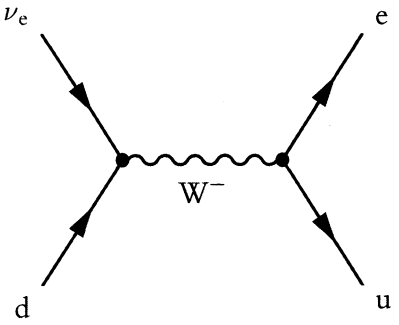


FIGURE 48 This is a weak scattering process in which the W couples to both leptons and quarks. As we know, quarks are confined within hadrons so this "pure" process can never be observed directly; but a simple twist of the diagram, with added spectator quarks, can represent the beta decay of the neutron. Neutron decay is mediated by the W particle. A d quark in the neutron converts to a u quark, thus changing the neutron to a proton, while the W decays into an electron and electron antineutrino.

The weak force mediators are extremely massive: the W^- and W^+ are 81.8 GeV, and the Z^0 is 92.6 GeV. *This* is the reason why the weak interaction seems so feeble. The uncertainty principle tells us that in order for two particles to be close enough to exchange a virtual particle with a mass of the order of 100 GeV, and thus participate in a weak interaction event, they have to be very close together — about 10^{-16} meters apart. This is unlikely to happen. (Though it is certainly not an impossible occurrence; on average, two quarks in a free neutron come close enough to exchange a W boson about once every ten minutes. But it is only by going to the high energies available in particle accelerators and forcing particles to close proximity that we can produce W and Z bosons in abundance.) The intrinsic strength of the weak interaction is not small at all. In fact, when experiments are performed at energies close the W and Z masses, the weak coupling constant α_W is measured to be $1/29$ — almost five times *larger* than the fine structure constant.

There are a couple of subtleties to the weak interaction that turn out to be important.

First, the natural assumption would be that the weak interaction converts quarks only within the same family, in the same way that it converts leptons only within the same family. Under this assumption the weak interaction would act on the (u, d) family and convert a u quark into a d quark, and vice versa. It would act similarly within the (c, s) and (t, b) families, but cross-family transitions would be ruled out. Well, this assumption is wrong. There are, for example, strangeness-changing weak processes; it has long been know that an s quark, say, can transmute into a u quark.

The way this is handled in the Standard Model involves one of those seemingly bizarre properties of quantum mechanics. As far as the weak

interaction is concerned, the quark families are not (u, d), (c, s), and (t, b); rather, they are (u, d′), (c, s′), and (t, b′). The primes indicate that the quarks are not the physical d, s, and b quarks; instead, they are quantum mechanical combinations of the physical d, s, and b quarks. So as far as the weak interaction is concerned, the d′ quark, for example, is mainly d, but there is a small amount of s and a trace of b. The same goes for s′ and b′. It sounds weird, but it makes perfect sense within the world of quantum mechanics. The amount of this "mixing" between families is an important number in the Standard Model. It would be good to be able to calculate it, but at present all we can do is deduce the number from experiment and then feed it back into the theory.

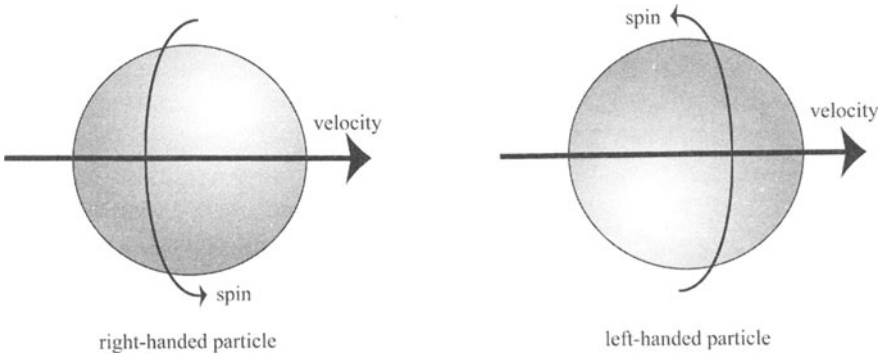


FIGURE 49 Suppose the two particles in this diagram are massless, and so move at the speed of light. The axis of spin for the particle on the left is in the same direction as the direction of propagation; it is a "right-handed" particle. The axis of spin for the particle on the right is in the opposite direction to the direction of propagation; it is a "left-handed" particle. If we could overtake the particle and look back at it, the handedness would appear to change; but since we can never catch up with a particle moving at light speed, the handedness for massless particles is conserved. Many experiments have shown that the weak interaction favors left-handed fermions (or right-handed antifermions).

Second, the notion of "handedness," or chirality, is important in weak processes. Consider a massless fermion — a spin-½ particle that travels at the speed of light. The spin can point in the same direction as the direction of propagation, in which case the fermion is "right-handed," or it can point against the direction of propagation, in which case it is "left-handed." Since we can never overtake a particle travelling at the speed of light, this handedness is a conserved quantity. For a long while, neutrinos were thought to be massless, and handedness applied very clearly to them: there were thus two "species" of neutrino — left-handed and right-handed. For massive fermions, such as the electron, the classification scheme is not so clean; since they do not travel at light speed we can imagine overtaking

an electron and seeing its handedness change as we look back at it. Quantum mechanical effects mean that an electron that starts out left-handed will evolve a component of right-handedness, and vice versa. When the electron moves at speeds very close to the speed of light, however, we can reasonably talk about a particle that possesses a particular handedness. We call these states of particular handedness the *chiral states* of a fermion.

Chirality

We met the word "chiral" before, in the chapter on symmetry. It comes from the Greek word meaning "hand," and it is used as a stem for many words in science. In botany, for example, the official name for the genus of fragrant wallflowers is *cheiranthus* — "hand flowers" — which comes from the custom of carrying these flowers in the hand as a bouquet. The stem appears in non-scientific contexts, too: chiromancy is divination by the hand — palm reading, in other words. In physics, chirality always has to do with whether something is left-handed or right-handed.

An important, if surprising, law of nature is this: *the weak bosons couple only to left-handed fermions* (or their right-handed antiparticles). Thus any realistic theory of the weak interaction must be chiral. As mentioned earlier, the discovery of this law came as something of a shock to physicists. They had assumed that nature obeys the laws of mirror symmetry: they thought that if you swap the space coordinates in mirror-image fashion then physics would be unchanged. That is indeed the case for the strong and electromagnetic interactions; but if you look through the mirror of the weak interaction, you find that nature treats left and right quite differently. It favors the left.

So the families of particles, as far as the weak interaction are concerned, are slightly different from what we might expect. The three lepton families that take part in weak interactions are $(e, \nu_e)_L$, $(\mu, \nu_\mu)_L$, and $(\tau, \nu_\tau)_L$, where the subscript denotes that these are left-handed states. The three quark families that take part in weak interactions are $(u, d')_L$, $(c, s')_L$, and $(t, b')_L$; again, the subscript denotes that these are left-handed quarks, and the primes denote that the quarks are mixtures of the physical states.

All this is rather involved, but it is the only way to make sense of the mountain of data on weak interactions collated by physicists. And chirality turns out to be of crucial importance: ensuring that our low-energy theories of nature are chiral provides a tight constraint on scientists who wish to develop unified theories of physics.

ELECTROWEAK UNIFICATION AND THE HIGGS

A union in partition.
William Shakespeare, *A Midsummer Night's Dream*

We understand the three fundamental interactions in terms of the exchange of bosons. This common underpinning of the theories leads to the hope of unifying the three interactions. In fact, the first step along this road was taken decades ago: the electromagnetic and weak interactions have been unified.

The electromagnetic and weak interactions seem to be two very different beasts. Electromagnetism couples solely to electric charge, whereas the weak force affects all particles, including the electrically neutral neutrinos. Electromagnetism acts over long distances, whereas the weak force is negligible over distances larger than an atomic nucleus. Electromagnetism does not distinguish between left-handed and right-handed particles, whereas the weak force is chiral. There are other differences, too. Nevertheless, physicists long ago recognized that there are deep theoretical links between the two interactions. In fact, as long ago as 1961 the American physicist Sheldon Glashow — building on even earlier work by his thesis advisor Julian Schwinger, one of the founders of QED — showed how we could consider the two forces to be merely different aspects of a single *electroweak* interaction.

Glashow started with the idea of local gauge symmetry, which had proved so successful for QED. To unify electromagnetism with the weak interaction, we need to identify the appropriate "charges" of the combined interaction which, when gauged, will lead us to a quantum field theory of the interaction.

Consider first, then, a left-handed doublet of particles upon which the weak interaction can act. For example, it could be $(e, \nu_e)_L$ — the electron and its neutrino. (We could equally choose any of the other five doublets.) We should be able to "swap" these particles at any spacetime point and find that the world is unchanged under this transformation, so long as a suitable gauge boson takes part in the process. This is similar to the way we can "swap" a red quark for a blue quark so long as a gluon takes part in the process.

The mathematics of a transformation in which two objects can be rotated into each other is described by the symmetry group SU(2). In this particular case, we can call the symmetry group $SU(2)_L$, where the subscript reminds us that we are dealing only with left-handed particles. The quantity being "rotated" here in the weak interaction is something called *weak isospin*. We can think of this quantity as being an arrow associated with a pair of particles. In the case of the $(e, \nu_e)_L$ doublet, if the arrow

The Group SU(2)

The group SU(2) can be used to describe the rotation of two objects into each other. For example, it can be used to describe a spin-½ particle in quantum mechanics: the spin can point either up or down the z axis, and thus SU(2) is appropriate. The same group can also be used to describe *isospin*, which was an important concept in the early days of the study of the strong interaction. The proton and neutron were thought to be two different isospin states of an entity called the *nucleon*; if the isospin was up then the nucleon was a proton, and if the isospin was down then the nucleon was a neutron. (This is the origin of the names of the up and down quarks.) Isospin has nothing to do with spin, but the same mathematical structure — described by SU(2) — is appropriate. For the weak interaction, we have *weak isospin*. (Weak isospin has little to do with either spin or isospin, but the reader will have gathered by now that physicists are often lazy when it comes to nomenclature.) The point is that the underlying mathematics are the same in all three cases, and the transformation on the left-handed doublet of particles can be described by the SU(2) group.

points up, say, then the particle might be an electron, while if it points down then the particle would be a neutrino. We are free to change the direction of the arrow from point to point, so long as a messenger particle — a gauge boson — conveys the information about the choice made.

We also need to consider a quantity called *weak hypercharge*, symbolized by Y, which is simply related to electric charge. The leptons and quarks all possess a value for the hypercharge; a right-handed tau has $Y = -2$, for example, whereas a right-handed neutrino has $Y = 0$. We can change the value of hypercharge from point to point, so long as a messenger particle — a gauge boson — conveys the information about the choice made. The symmetry group of a phase rotation of hypercharge is $U(1)_Y$, where the subscript reminds us that we are not talking about the usual U(1) of electromagnetism. Again, the mathematics of charge and hypercharge is the same, but the physical meaning is slightly different.

Glashow studied a local gauge theory based on the product of these two groups: $SU(2)_L \times U(1)_Y$. As mentioned earlier, a local gauge symmetry demands the existence of massless gauge bosons. The SU(2) is responsible for three such bosons — call them W^+, W^-, and W^0. The U(1) is responsible for another — call it B^0. The rules of the quantum mechanical world allow the W^0 and B^0 bosons to mix to form two different massless bosons — call them Z^0 and γ. (The amount of mixing is determined by a number

called the "weak mixing angle" — which is another of those important numbers in the Standard Model that must be measured experimentally and fed back into the theory.) So the $SU(2)_L \times U(1)_Y$ theory has four massless gauge bosons called W^+, W^-, Z^0, and γ. But it is not just that we can *call* these bosons by familiar names; for this particular choice of symmetry — local symmetry under weak isospin and weak hypercharge — we really can identify the four gauge bosons as the three weak interaction bosons and the photon. In other words, the $SU(2)_L \times U(1)_Y$ symmetry provides us with an integrated theory of electromagnetic and weak interactions! Because there are two separate groups there are two different coupling constants corresponding to the two different interactions; so there is not complete unification in terms of coupling strengths. But we do have a unified framework for describing the two interactions; we have a single interaction that manifests itself in two ways.

This was perhaps the most important advance in the path towards the unification of forces since 1865, when James Clerk Maxwell began publishing work showing that electricity and magnetism are merely different aspects of a single electromagnetic force. Or at least it would have been, were it not for a glaring problem with Glashow's idea: all the gauge bosons in his theory are massless. The photon is indeed massless, but Glashow himself knew that the other three mediators had to be extremely massive in order to explain the short range of the weak interaction. Glashow had no explanation for why the photon alone should remain massless; he simply remarked that "it is a stumbling block we must overlook."

The underlying trouble with Glashow's idea is symmetry — there is too much of it. The $SU(2)_L \times U(1)_Y$ symmetry, if it holds exactly, demands that all the bosons have zero mass. This does not reflect the world as we see it. But in 1967, Abdus Salam and Steven Weinberg independently showed how the theory could retain its beautiful symmetry at a deep level and still give rise to massive weak gauge bosons. Their idea was that the symmetry of the theory is broken spontaneously — "spontaneously" meaning that the symmetry breaking occurs naturally.

There is nothing mysterious about spontaneous symmetry breaking. There are many examples in physics. Hold a drinking straw between the palms of your hands and you have a physical system that can be described by equations possessing rotational symmetry. Press your palms together and the straw bends. The symmetry is broken. You cannot necessarily predict how the straw will bend; it could bend "up," "down," "sideways," or in any other direction. This unsymmetrical situation, however, is the stable solution of perfectly symmetrical equations. Or consider an example we looked at in an earlier chapter: water. A pail of water has a high level of rotational

FIGURE 50 Abdus Salam (left) and Steven Weinberg showed how the mechanism of spontaneous symmetry breaking could give mass to the weak gauge bosons whilst keeping the photon massless. They showed how Glashow's proposal for electroweak unification could be made to work. All three men shared the 1979 Nobel prize for physics.

symmetry — it looks the same from whatever angle we view it. But if we cool the water so that it freezes — if, in the jargon of physics, it undergoes a *phase transition* — then this rotational symmetry is broken, because ice has a crystalline structure. Ice looks different when viewed from different angles. This is an important point: whenever a physical system undergoes a phase transition, the symmetries of the system can be broken. Perhaps the best example of spontaneous symmetry breaking involves magnetism. A ferromagnet is just a collection of interacting ions, tiny little magnets that point in all directions. The equation giving the total energy is rotationally invariant; the magnetic domains can point in any direction and so there is no net magnetic field. But it turns out that the lowest energy state of the ferromagnet has all the magnetic domains pointing in the same direction: rotational symmetry is spontaneously broken. If you heat the ferromagnet above some critical temperature, so that the magnetic domains jiggle about, then the ferromagnet undergoes a phase transition; it loses its magnetism and symmetry is restored. Cool it again, below the critical temperature, and the symmetry breaks; all the domains point in one direction (the precise direction is random, an accident of history) and the ferromagnet is once more magnetic.

Salam and Weinberg introduced the *Higgs field* into the Standard Model as a mechanism for inducing spontaneous symmetry breaking. (The field is named after the British physicist Peter Higgs, who was among the first to work on the problems of symmetry breaking in physics. Other people who studied similar details of symmetry breaking in field theory around

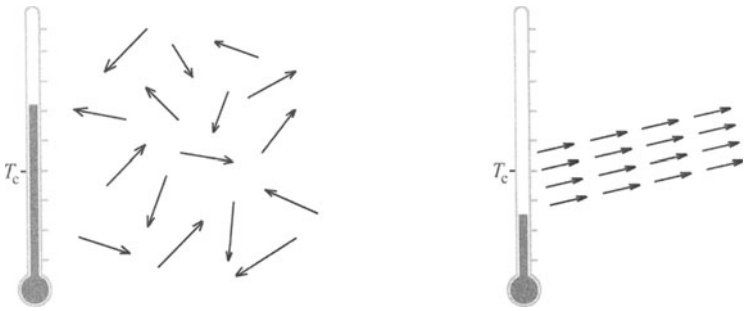


FIGURE 51 Left: Above a critical temperature, the magnetic domains in a ferromagnet all point in different directions. In this rotationally symmetrical system there is no net magnetic field. Right: Once the temperature drops below the critical value, the magnetic domains all point in the same direction. The rotational symmetry is broken and a magnetic field appears. The particular direction of the field is random: suppose you reheat the ferromagnet above the critical temperature, so that the magnetic field disappears, and then cool it again — the magnetic field will point in some other direction.

the same time include Robert Brout and François Englert in Belgium, and, even earlier, Philip Anderson in America.)

The Higgs field, if it exists, permeates all of space and couples to all fundamental particles. It is the coupling of the Higgs field to the gauge bosons of Glashow's electroweak theory that gives the W and Z particles their masses. Recall from our earlier discussions that a massless spin-1 boson can have only *two* degrees of freedom, the two transverse polarizations. A *massive* spin-1 boson has *three* degrees of freedom: a longitudinal polarization in addition to the two transverse polarizations. The Higgs mechanism proposes that there is a field with *four* physical polarizations permeating space. When the W^+ interacts with the field it "eats" one of the polarizations and become massive as a result; the W^- "eats" another of the polarizations and also becomes massive; the Z^0 "eats" a third and it too becomes massive. The photon, which declines to partake in this feast, remains massless. And the original field? This is left with just one polarization — it is a scalar boson (that is, a spin-0 boson). This is the so-called Higgs boson (symbol h).

This is a subtle concept, so it bears repeating. The electroweak theory has an $SU(2)_L \times U(1)_Y$ symmetry, with four massless gauge bosons. When these bosons interact with the Higgs field, a field that fills our Universe, the symmetry breaks. Three of the bosons (W^+, W^-, and Z^0) acquire mass; the fourth boson (γ) remains massless; and there is a massive scalar boson (h) left over.

It turns out that the Higgs mechanism can give masses to all the fundamental fermions. As a quark or lepton moves through space, it interacts

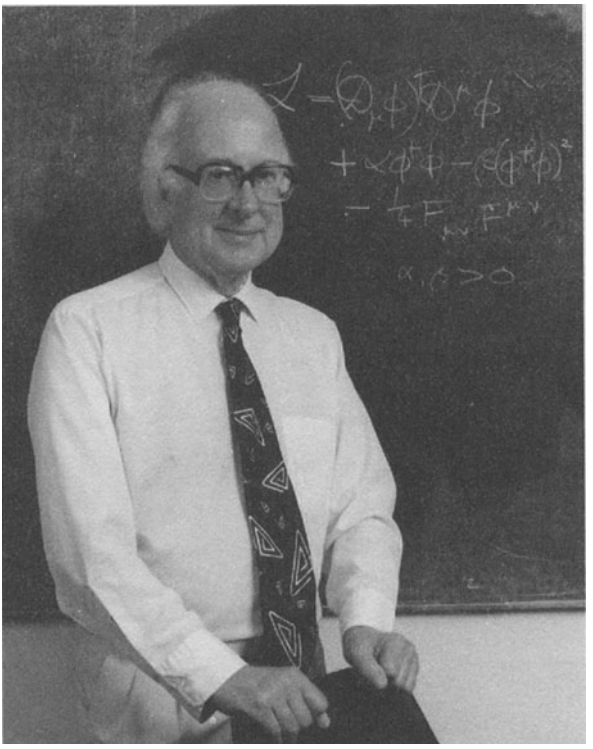


FIGURE 52 Peter Higgs is an emeritus professor of theoretical physics at Edinburgh University. He has won many honors for his work on field theory including, with Robert Brout and François Englert, the 2004 Wolf prize. (After the Nobel prize, the Wolf prize is often thought to be the most prestigious award in physics.)

with the Higgs field; the field becomes distorted in the vicinity of the particle. It is this distortion that causes the particle to have mass. (Figure 53 is a cartoon illustration of this mechanism, inspired by David Miller.) It is sometimes said, therefore, that the fact that matter possesses mass can be explained by the Higgs mechanism. Things are slightly more complicated than this saying suggests: as we have seen, the mass of the proton, say, is due mainly to the energy that binds the quarks together rather than the sum of the quark masses. But the fundamental particle masses are due to the Higgs mechanism, so in this sense the saying is correct.

So by 1967, physicists had a unified electroweak theory. In 1971, the Dutch theoretician Gerardus 't Hooft showed that the theory was renormalizable. Just as with QED, the infinities can be made to go away, and the theory gives you well-defined answers to your questions. The electroweak theory was here to stay.

The success of the electroweak theory, combined with the success of QCD, meant that the Standard Model was quickly established as our best theory of the microworld. The Standard Model of course actually consists of *two* distinct theories, but both of them are based on the ideas of local

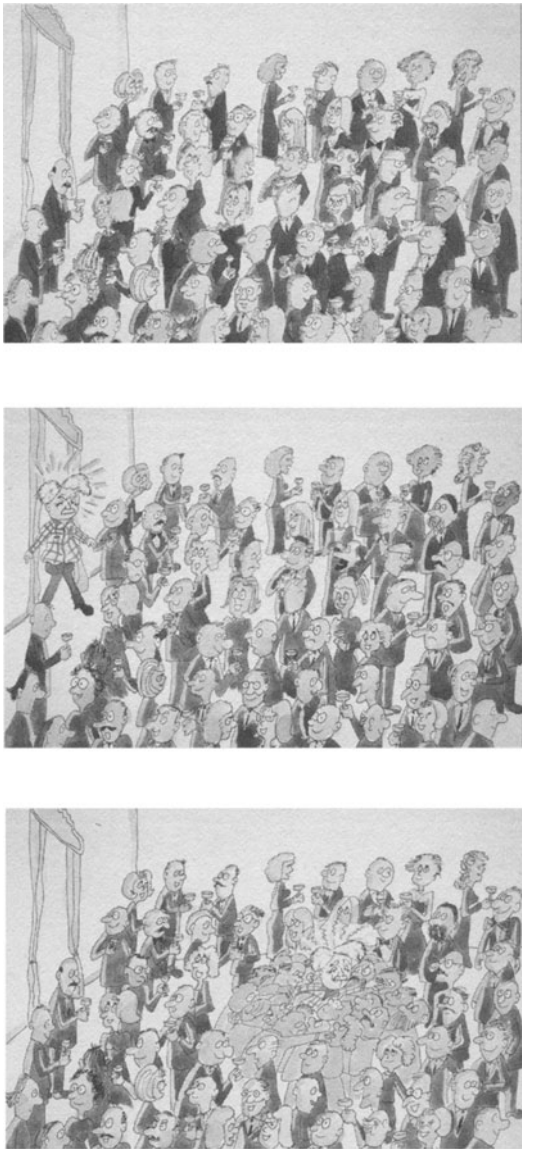


FIGURE 53 In 1993, the British Minister for Science challenged physicists to describe the Higgs mechanism on one sheet of A4 paper. David Miller won with a political analogy; the cartoon here is a physics version. Physicists at a social gathering are uniformly spread throughout a room, happily chatting to their nearest neighbor. Einstein enters and tries to walk across the room. Physicists in his vicinity, eager to talk to their hero, crowd around him and impede his progress; only after he has passed do they return to their original spacing. It is as if he has gained mass. (In Miller's version, Margaret Thatcher visits a cocktail party thrown by Conservative Party members.) Analogously, fundamental particles interact with the Higgs field that permeates space; the local distortions of the Higgs field gives rise to the particle masses.

gauge symmetry. The strong interaction, described by QCD, is based upon the gauge group $SU(3)_C$, while the electroweak interaction is based upon $SU(2)_L \times U(1)_Y$. Thus many physicists often refer to the Standard Model simply as $SU(3) \times SU(2) \times U(1)$.

The achievement of Glashow, Salam, and Weinberg is on a par with Maxwell's unification of the electric and magnetic forces. The achievement is at the cost of introducing one more fundamental particle into the Standard Model: the scalar Higgs boson. At the time of writing, the Higgs boson has not yet been observed in experiment. It is a missing link, the only piece of the Standard Model that awaits experimental verification.

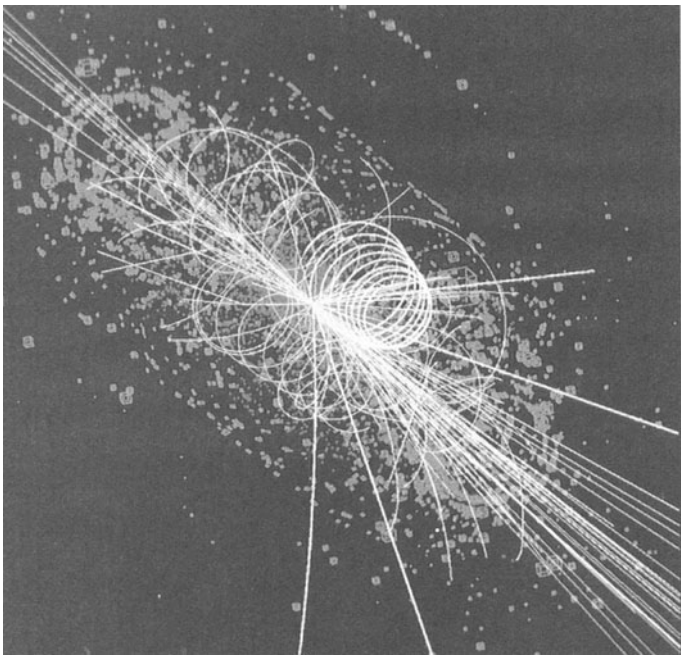


FIGURE 54 The Higgs particle has not yet been observed, but by now physicists have a good idea of where they should look to find it. This computer simulation shows what Higgs production might look like at the Large Hadron Collider. Here two high-energy protons have collided to produce a Higgs boson, which then decays into four muons. Such an event might be seen in the CMS detector.

The Mass of the Higgs

There are tantalizing hints that the LEP machine at CERN saw the Higgs in 2000, but this has not been confirmed. We do know, from LEP and similar machines, that the mass of the Higgs must be greater than 114 GeV. We also know that the Higgs cannot be more massive than about 250 GeV. When it comes on line in 2007, the Large Hadron Collider will certainly have the energy to discover the Higgs boson — if the particle exists.

Our view of the Universe is distorted, made ugly because we live in a low-energy world. At these low energies we see two distinct forces: electromagnetism and the weak interaction. But the Universe was not always like this. In the first instants after the Big Bang, the temperature of the Universe was much higher than it is now. If we follow the behavior of the electromagnetic and weak interactions backwards in time, we find that the Universe underwent a phase transition. Before the phase transition occurred, roughly 10^{-11} seconds after the Big Bang, the Universe possessed the full, beautiful symmetry of electroweak theory; the electromagnetic and weak interactions melded into the single electroweak interaction. Nowadays we can see the symmetry only by doing experiments at particle colliders. This is perhaps our first glimpse out of the "cave": the notion that the Universe started out in a much simpler, more symmetric state. It is only now the Universe is old and cold that the symmetry is hidden. Physicists have doggedly followed this idea and the trail has led them to the wild ideas we will soon be in a position to discuss.

THE SUCCESSES OF THE STANDARD MODEL

One thing is forever good;
that one thing is success.
Ralph Waldo Emerson, *Fate*

I referred earlier to the Standard Model of particle physics as being "our best theory." Here are just five reasons why.

First, with one exception, all the fundamental fields of the Standard Model have been discovered and their properties measured. The only exception is the Higgs boson, and if it exists the LHC will find it in a few years' time. (If it turns out that the Higgs does not exist, that would not necessarily be fatal for the Standard Model. There may be other ways of breaking the electroweak symmetry and giving mass to the particles. In this case, however, the usual formulation of the theory would certainly need to be revised.) Several of the fundamental particles were *predicted* by the Standard Model and their properties described. For example, physicists knew where to look for the top quark; many of its properties were known from theory several years before it was found in experiments. Conversely, no one has yet discovered a fundamental field that does not belong within the Standard Model.

Second, the menagerie of subatomic "particles" that so baffled physicists in the 1950s and 1960s is no longer confusing. These objects are not the fundamental particles they were once thought to be; they are just

bound states of the truly fundamental entities that combine in ways allowed by the Standard Model. And no subatomic object has been discovered with properties that are forbidden by the Standard Model. No one has found a baryon with strangeness 5, for example, or a meson with charm 3.

Other Particles in the Standard Model

It is possible that the Standard Model — or minimal extensions of the model — might contain particles that have not yet been observed. One example is the *axion*.

A remaining puzzle in QCD is why the theory is observed to obey a symmetry known as CP. This symmetry designates what the world looks like if we exchange particles with antiparticles (a transformation known as charge conjugation, symbol C) and simultaneously exchange left with right (a transformation that essentially means we look at physical systems in a mirror; the transformation is known as parity, symbol P). The weak interaction violates CP symmetry, and there are good reasons to expect the strong interaction to violate it also. But experiments show that the strong interaction obeys CP symmetry.

In 1978, Helen Quinn and Robert Peccei introduced a mathematical symmetry, called axial symmetry, into the Standard Model. Such a symmetry would naturally suppress the CP-violating effects present in the theory of strong interactions. The axial symmetry must be spontaneously broken, however, and the result is the appearance of a new spin-0 boson called the axion, just as the spin-0 Higgs boson appears when the electroweak symmetry breaks spontaneously. (The axion continues the long tradition of light-hearted names for particles. It was named after a brand of laundry detergent.) The axion, if it exists, would have very little mass — possibly much less mass even than neutrinos — and would interact as weakly as neutrinos. Several experiments have searched for the axion, but so far without success.

If the axion is found it will fit nicely into the existing framework of the Standard Model. However, recent work suggests that the strong-CP problem can be solved without recourse to the axion.

Third, one of the goals of particle physics has been to calculate quantities like cross sections (the likelihood of particles scattering off each other) and decay rates (the typical length of time that a particle can exist before it converts into a less massive particle). The electromagnetic sector of the Standard Model has seen spectacularly successful in comparing the results

of theory and experiment. As we saw on page 82, QED calculations involving the motion of an electron in a magnetic field are correct to ten decimal places. When theory and experiment agree so closely, we must be doing something right. Calculations in the QCD sector of the Standard Model are more difficult, and the experiments more messy, but there is still a good level of agreement between theory and experiment.

Fourth, there have been *trillions* of events recorded in particle accelerators over the past half century. This is a huge amount of data. *None* of this data contradicts the Standard Model. (In March 2004, a team of physicists at the Brookhaven National Laboratory claimed that they might be seeing effects of physics beyond the Standard Model. They sifted through 7.8 trillion events looking for rare decays of a particle called a K meson. The Standard Model suggests that the decay of a K meson into a pion and a neutrino–antineutrino pair should occur once in every 13 billion decays. But the team saw three of these rare events, suggesting that the decay rate is once in every 7 billion decays. This might be a signal of physics beyond the Standard Model — but as with other such claims (see page 82) the experiment must run for much longer before the issue is settled. Three events may represent just a statistical fluke.)

Fifth, the Standard Model is ultimately based upon just a few basic principles: quantum mechanics and special relativity combine to form quantum field theories; the theories possess local gauge symmetries (which are spontaneously broken in the electroweak sector); the theories are renormalizable. From these basic principles, the Standard Model flows. This makes it difficult to tinker with the Standard Model, except in restricted and well understood situations; the model *must* have the structure it does because the principles upon which it is based can combine in only certain ways.

It would seem that, armed with knowledge of the Standard Model, we are well on the way to a final theory — a theory that could explain everything we want to know about how to construct the fundamental aspects of our Universe. Well, the Standard Model is good — but not *that* good. In fact, when you think about the Standard Model, it does not take long before you realize that it has to be replaced with something better.

The Standard Model Is Not A Final Theory

It is a condition which confronts us — not a theory.
Grover Cleveland, *Annual Message, 1887*

At first glance, the number of truly fundamental particles is satisfyingly small: a family of leptons (e, ν_e) and quarks (u, d) that, for some reason,

is replicated twice over. And the fundamental interactions between these particles are well understood by QCD and electroweak theory, both of which are based on beautiful ideas of gauge symmetry. This all seems elegant enough to form the basis of a final theory. But is it really so elegant?

You can reasonably argue that there are an embarrassingly *large* number of fundamental particles. There are six leptons, all of which have associated antiparticles, making 12 leptons in all. Then there are six types of quark, each of which comes in three colors, making 18 quarks; counting the antiquarks gives us a total of 36 quarks. Now consider the gauge particles that mediate the interactions: 1 photon, 3 weak bosons, 8 gluons — so 12 in total (the bosons serve as their own antiparticles). And there is at least one Higgs boson (in some formulations of the theory, there could be more). Looking at it this way, there are at least 61 fundamental particles in the Standard Model. That seems an uncomfortably large number for a final theory of physics.

Much worse than the *number* of particles is the fact that the *masses* of all these particles have to be put into the Standard Model "by hand." In other words, we have to *measure* their values experimentally and then feed them in to our theories. The Standard Model tells us nothing about what these masses should be. It would not be so bad if there were some obvious pattern to the fermion masses, but there is none that can be discerned. The t quark is extremely massive (about 180 GeV), the electron is light (0.511 MeV), and the neutrinos are extremely light. (Figure 55 gives a visual representation of the different particle masses in the Standard Model.) We believe we know where the masses come from: it is the coupling of the fermion fields to the Higgs that gives the various fermions their mass. But that does not tell us what the masses should be, because we cannot calculate the strength of these couplings. Since the Higgs boson must have a mass of between about 114–250 GeV, we can argue that the "natural" coupling to the Higgs field is possessed by the t quark. But then we have to ask why the couplings of the other fermions are so small. The Standard Model remains silent on such questions.

Various other parameters, such as the coupling constants and the various mixing angles, must be put into the theory "by hand." A final theory should constrain some, ideally perhaps all, of these parameters.

There are many other questions, both general and specific, for which the Standard Model has no answer. To give just one example: why are there *three* families of otherwise identical particles? The Bruce Hornsby song says "that's just the way it is," and such an answer suffices in many walks of life. But that is not a satisfactory answer from a final theory of physics.

There are problems on aesthetic grounds, too. The symmetries of the Standard Model are described by classical Lie groups, like SU(2) and SU(3).

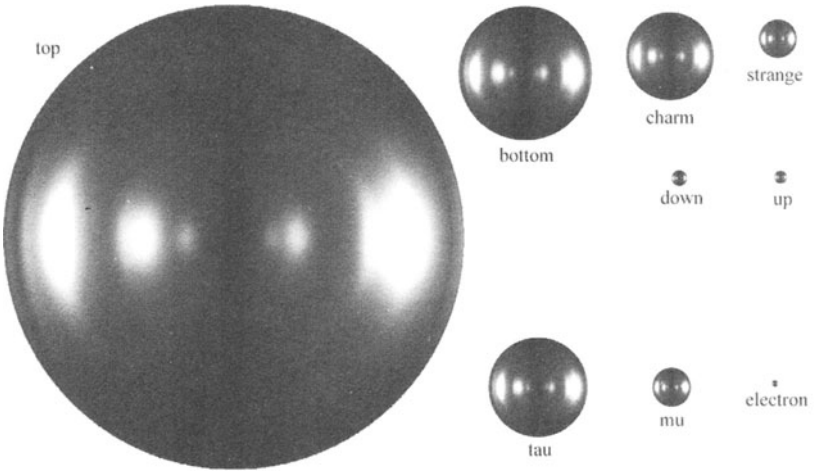


FIGURE 55 The rest masses of the fundamental particles are here represented by the radius of a sphere. (On this scale, the neutrino masses are far too small to see.) The various particle masses span a wide range, and there appears to be no obvious pattern to the masses. Only the top quark has a mass that is natural in the theory; explaining the small value of all the other masses is a major challenge for physicists.

Unlike the exceptional Lie groups, which are tightly constrained, there are an infinite number of groups like SU(N). Mathematically, the number N can take any value. Why should nature choose particular values of N — why choose SU(2) and SU(3), for example, rather than say SU(475)? More importantly, the Standard Model does not provide a unified theory of particle interactions. There are *two* distinct sectors: SU(3) describes the QCD sector and SU(2) $\times$ U(1) describes the electroweak sector. Putting them together as SU(3) $\times$ SU(2) $\times$ U(1) is like trying to fit two mismatching pieces of a jigsaw puzzle together by banging them together with your fist. On the other hand, the underlying concepts and mathematics of these gauge theories are essentially the same. The Standard Model almost *demands* further unification.

How can the Standard Model be so successful if it is not, somehow, "the Truth?" Well, the situation has happened before in physics. For decades physicists used a theory of the weak interaction that was extremely successful in terms of its agreement with experiment, but even the author of the theory — Enrico Fermi — knew that ultimately it had to be wrong. Everyone knew the theory was wrong because it was non-renormalizable: at high energies the theory gave infinite answers. The source of the prob-

lem was that Fermi's theory considered weak interactions to act at a point, whereas we now know that the weak interaction is spread out since it takes place through the exchange of W and Z particles. At the energies available at colliders in Fermi's time, though, which were well below the mass of the W and Z, the interaction *is* effectively point-like. The Fermi theory is thus an excellent low-energy *effective* theory of the weak interaction. Perhaps the same is true of the Standard Model: perhaps it is a low-energy effective theory of some more general theory that holds at higher energies. Indeed, perhaps we will have to develop several different effective theories, each valid in a particular energy regime, on our path to a unified theory.

To many physicists, then, the Standard Model — although hugely successful — is simply too arbitrary, too ugly, too *forced* for it to be the final theory we are looking for. And — the killer blow — in the final instance it *cannot* be a final theory: it says nothing about gravity.

Even before the Standard Model was fully accepted, many physicists were already looking for theories that might lie beyond it.

5

GUTS AND SUSY

Seagoon: Gad, Bloodnok, I admire your guts.
Bloodnok: Why, are they showing?
Spike Milligan, *The Booted Gorilla*

The twin pillars of physics rest on solid ground. In their respective domains general relativity and the Standard Model have passed all experimental tests. Anything beyond these pillars is — to a lesser or greater extent — speculation. Any claim made for physics beyond these pillars has to be taken with a grain (and often a bushel) of salt. As mapmakers of old used to write on the unexplored regions of their charts: here be dragons!

Moving beyond general relativity has proved to be a difficult task. But physicists have long hoped to obtain a more natural theory of the fundamental particles and their interactions. The hope is well-founded, since it seems that some of the tricks used to develop the Standard Model can be recycled and then used to improve upon it. The first step into the speculative reaches of physics, then, is the investigation of ways of extending the Standard Model. The two most venerable approaches are grand unified theories (GUTs) and supersymmetry (SUSY). Let's begin by looking at grand unified theories.

GRAND UNIFIED THEORIES

Everything that is unknown is taken to be grand.
Tacitus, *Agricola*

If we want to move beyond the Standard Model, our first step would seem to be obvious. We know how to unify the weak and electromagnetic interactions: we write down a theory with a large enough gauge symmetry to accommodate the four mediators. Then, through the action of the Higgs mechanism, we break the symmetry so that three of the mediators become massive while the fourth (the photon) remains massless. Surely what we need to do is to repeat the trick: pick some "grand unified" gauge symmetry that is large enough to accommodate gluons in addition to the electroweak bosons — in other words, a group large enough to contain $SU(3) \times SU(2) \times U(1)$ as a subgroup — and use the Higgs mechanism to break the symmetry. We would then have a grand unified theory that would have all of particle physics in its compass.

In this picture, the Universe was once much simpler and more symmetrical than it is now. At some extremely high energy (such as would occur in the Universe at a fraction of a second after the Big Bang) the three particle interactions — strong, weak, and electromagnetism — would be truly unified. Only as the Universe cooled and underwent phase transitions would the symmetries break and leave us with the distinct forces we observe today.

Many groups are large enough to contain the $SU(3) \times SU(2) \times U(1)$ symmetry of the Standard Model as a subgroup. So there are many possibilities for the grand unified symmetry. The simplest group that will do the job is SU(5). This is the group that Sheldon Glashow and Howard Georgi used in 1974 to construct the first grand unified theory.

Physicists simply had to look in the group theory books to find the properties of the SU(5) group. The books told them that a grand unified field based on SU(5) must contain 24 gauge bosons, 12 of which would correspond to the familiar force mediators (the eight gluons, the three weak bosons, and the photon). The other 12 components would correspond to completely *new* fields. The quanta associated with these new fields — new force mediators — are often called X and Y bosons. In the SU(5) model there are three X bosons, each with electric charge $-4/3$ and a color charge (X_r, X_g, X_b); there are three Y bosons, each with electric charge $-1/3$ and a color charge (Y_r, Y_g, Y_b); and each of the X and Y bosons has an antiparticle, making 12 bosons in total.

In the SU(5) model, the quarks and leptons are grouped together in particular ways. For example, the d quark (which, as we know, comes in three colors) forms an extended family of particles along with the antielec-

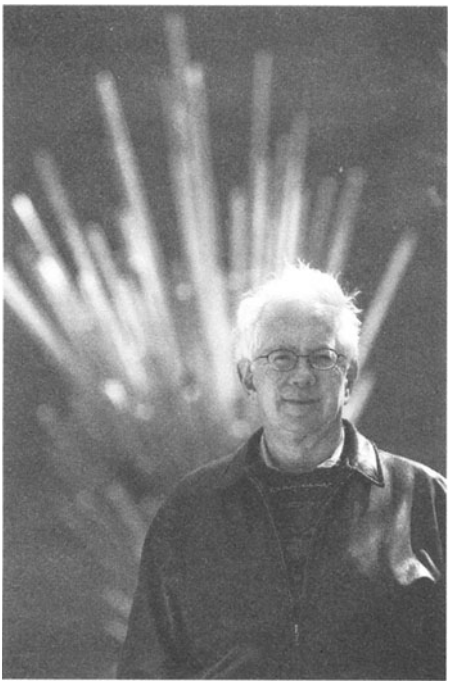


FIGURE 56 In 1979 Sheldon Glashow, then at Harvard University, shared the Nobel Prize with Abdus Salam and Steven Weinberg for unifying electromagnetism with the weak unification. He was also a pioneer in the development of grand unified ideas.

tron $\bar{e}$ and the electron antineutrino $\bar{\nu}_e$; thus we have a family of particles (d_r, d_g, d_b, $\bar{e}$, $\bar{\nu}_e$) that in a sense belong together in the SU(5) grand unified theory.

Interestingly, in a grand unified theory such as SU(5), the sum of the electric charges of an extended family of particles must be zero. For the family of particles just mentioned, the total electric charge is zero only if the d quark has charge $-1/3$. The factor 3 comes from the number of colors: the sum of the electric charges of d_r, d_g, and d_b is -1, which balances the $+1$ charge of the antielectron. (The neutrino, of course, is neutral.) This balance only occurs if there are three differently colored quarks in the family. In this model, quarks must have electric charges in units of $1/3$ because they come in three different colors! The electrical neutrality of matter, which is fundamental to our own existence and which arises because the charge on the electron is equal and opposite to the charge on the proton, has no explanation in the Standard Model. It is simply one of those coincidences we must take on faith. We now see that it can be explained as a consequence of a grand unified theory. This is a beautiful result.

The SU(5) model has other successes to its credit. For example, it predicts the ratios of the three coupling constants. And it seems to make a reasonable prediction of the value of the weak mixing angle, the value

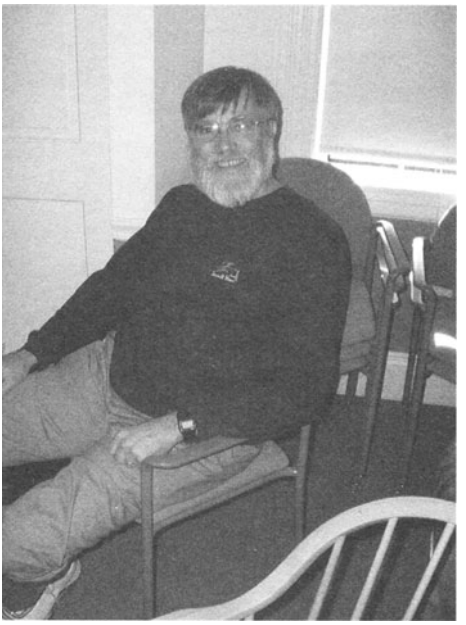


FIGURE 57 Howard Georgi, of Harvard University, helped develop the first grand unified theories. He continues to produce exciting new ideas in the field.

of which is a free parameter in the Standard Model. In the SU(5) grand unified model we have a theory that *tells* us the value that this fundamental parameter must take.

All this sounds promising, but is there any way of putting GUT ideas to experimental verification, as physicists have done so successfully with the Standard Model?

A grand unified symmetry such as SU(5) implies that we can, at each point in space and time, "rotate" one member of an extended family into another member of the family and find that the world is unchanged — so long as a gauge boson carries information on the choice that has been made. The key property of the X and Y bosons, therefore, is that they mediate interactions between particles in the same extended family; in particular, they can convert a quark into a lepton. Perhaps this property of GUTs, a feature that is absent in the Standard Model, might be observable? There are many possible processes involving X and Y bosons; Figure 58 shows a typical process.

In Figure 58, three quarks enter (uud) but only two quarks leave (u$\bar{u}$), along with an antielectron. The exchange of an X boson has changed a quark into a lepton. In this case the d quark has changed into an anti-

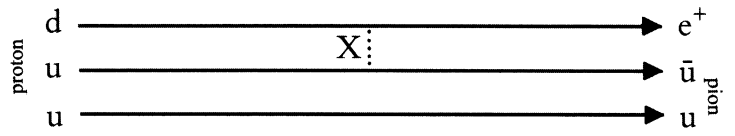


FIGURE 58 Proton decay is allowed in grand unified theories. In the process shown here, the decay is via a pion and antielectron: d and u quarks in a proton exchange an X boson and convert into an antielectron and antiup quark. Many other decay channels are possible.

electron, a reaction that is possible because these particles belong to the same extended family of particles in the SU(5) model. We thus glimpse a new and beautiful symmetry in GUTs: the distinction between quarks and leptons, which is a feature of the Standard Model, begins to break down. At a very deep level, the fermions — the "particles of matter" — lose the ugly division between quark and lepton.

The process described by Figure 58 may have an underlying beauty, but it is also rather disturbing. It implies that the proton can decay into an antielectron and a neutral pion. In fact, many other channels would be available for proton decay. Proton decay does not happen in the Standard Model, because there is a law of total-quark-number conservation, which leads directly to the law of baryon-number conservation. Since the proton is the lightest baryon, the proton is stable. This is just as well, since the stability of all the matter we see around us depends upon the stability of the proton. According to grand unified theories, however, the proton is unstable. Diamonds are not forever. Eventually, all protons will decay, atoms will disintegrate, and ultimately we will be left with a soup of electrons and neutrinos.

The detection of proton decay would be a clear experimental test of grand unified theories. Before considering this further we have to counter the obvious fact that, although GUTs may predict proton decay, the proton certainly *seems* to be stable. Indeed, some simple arguments immediately indicate that the average proton lifespan is at least 10^{16} years (see the boxed text on page 118). The present age of the Universe is about 10^{10} years — so 10^{16} years is a long time indeed. If proton decay is allowed in grand unified theories, why would it be such a rare phenomenon?

The answer lies in the *extremely* large mass of the X and Y bosons. We can get an idea of the masses of these particles by looking at the energy at which grand unification is likely to occur. According to renormalization ideas, the strength of an interaction depends on the energy at which we measure it. Coupling constants are not constant at all. There are three coupling constants in the Standard Model, two of which belong to the electroweak sector. One of these electroweak coupling constants starts out small

Human and Proton Lifespans

Our very existence constrains the half-life of proton decay: the human body contains about 10^{28} protons, so if the proton lifetime were less than 10^{16} years an average human body would produce about 30,000 proton decays per second. Your body would be a radiation health hazard! From the simple fact that you are reading this book, you can deduce that the proton must have a half-life of more than 10^{16} years.

and then *increases* very slowly with energy; the other starts out slightly larger and then *decreases* very slowly with energy. Clearly, there must be some energy at which they have the same value. The third coupling constant in the Standard Model belongs to the strong sector; at the electroweak scale it is much larger than the other coupling constants, but it decreases rather quickly with increasing energy. If the three coupling constants had the same value at the same energy, then grand unification would have taken place. In 1974, a trio of Harvard-based physicists — Georgi, Weinberg, and Helen Quinn — extrapolated the three coupling constants to extremely high energies within the SU(5) model. As shown in Figure 60, the coupling constants did indeed seem to converge on the same value. According to their studies, the grand unification energy was about 10^{15} GeV — not too far away from the Planck energy of 10^{19} GeV, at which point a quantum theory of gravity is required.

FIGURE 59 Helen Quinn's paper, authored with Howard Georgi and Steven Weinberg, was the first to discuss how the different coupling constants of the strong, weak, and electromagnetic interactions can merge to a single coupling constant in a grand unified theory. In a paper authored with Roberto Peccei, she postulated the existence of a near symmetry of the universe to explain how strong interactions can maintain the symmetry between particles and antiparticles while weak interactions do not.

The grand unification energy is truly huge — about the same amount of energy as the kinetic energy of a typical family car travelling at 50 miles per hour. That may not seem large, but an automobile contains about 10^{30}

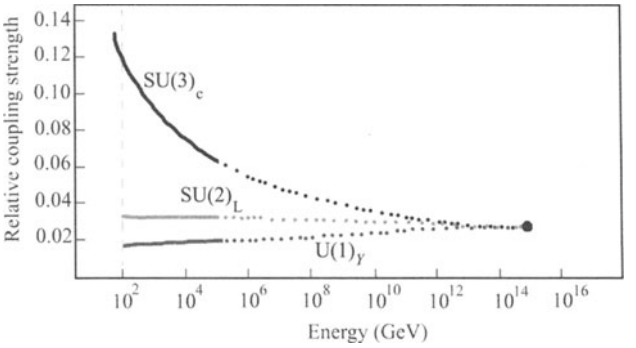


FIGURE 60 This shows the variation with energy of the three coupling constants of the Standard Model within the SU(5) GUT framework. Below 100 GeV, the electroweak interaction dissociates into electromagnetism and the weak nuclear force. But as energy increases, the two electroweak coupling constants appear to converge — and they converge with the color force coupling constant at a common energy: roughly 10^{15} GeV. This is the GUT energy.

particles. To achieve the grand unification energy, we would have to put *all* of the energy into just one subatomic particle! *This* is the reason why we do not observe proton decay. The uncertainty principle tells us that two quarks would have to be less than 10^{-31} meters apart before they would be close enough to exchange a virtual X boson. This is 10^{15} times smaller than the size of a proton. At the low energies of the everyday world — even at the high energies of our particle colliders — there is simply a negligible chance of the quarks in a proton exchanging an X or Y boson. If you wait long enough, then quantum physics tells you that anything that *can* happen *will* happen. After all, the quarks in a neutron sometimes get close enough to exchange a W boson and cause the neutron to decay into a proton. It happens every few minutes. But the W boson is less than 10^2 GeV in mass. It will take a long time for a particle of mass 10^{15} GeV to be exchanged by quarks and cause a proton to decay. How long? Well, depending upon the details of the calculations, the simplest grand unified theory based on SU(5) symmetry puts the proton lifetime at just over 10^{28} years, though this figure can rise by an order of magnitude or two, depending upon the details of the calculation.

This finding, that the GUT scale is about 10^{15} GeV, was dramatic. The scale was large enough to explain why proton decay had not been observed, and it was intriguingly close to the Planck scale, suggesting that a link with gravity was not far off. Furthermore, the GUT scale was not *beyond* the Planck scale, which would have rendered the whole scheme implausible.

It is not surprising that physicists were excited by the SU(5) GUT model. In addition to the points we have already discussed, the model had

important applications for cosmology (see the boxed material on this page and on page 123). By the early 1980s, grand unified theories had become the subject of intense theoretical investigation, but any experimental backing for GUTs was lacking. In particular, the non-observation of proton decay remained a problem for proponents of grand unification.

Grand Unification and the Existence of Matter

The Big Bang should have created particles and antiparticles in equal numbers. When particles meet their antiparticles, however, they mutually annihilate leaving behind only energy (in the form of photons). If there were initially equal numbers of particles and antiparticles, the Universe would now be filled only with photons. Since we see matter all around us, some mechanism must have produced more matter particles than antiparticles. We can measure the asymmetry between matter and antimatter by counting the number of cosmic background photons that remain. If we do that we find roughly one particle of matter for every 10^{10} photons. In other words, in the very early history of the Universe, for every 10,000,000,000 antimatter particles there were 10,000,000,001 matter particles. The 10 billion matter–antimatter pairs annihilated, leaving behind one particle of matter and billions of photons.

But how can we explain this asymmetry between particle and antiparticle, when the laws of physics seem to treat them identically? In 1967, the Russian physicist Andrei Sakharov showed that three conditions had to be met before a Universe containing initially equal amounts of matter and antimatter could evolve into a matter-dominated Universe. The first of these conditions was that the proton had to decay. So GUTs can provide part of the reason for why we see matter in the Universe — and, indeed, for why any of us are here at all!

Given the likely lifetime of the proton, how can we observe proton decay? Clearly, we cannot watch one proton for 10^{28} years in the hope of seeing it decay. But we *can* watch 10^{28} protons for one year in the hope of seeing one of them decay. And the more protons we can gather in one place, the more chance we have of seeing a decay. Well, a cube of water 10 meters on each side contains about 10^{33} protons; if we shield the water from stray radiation, and surround it by particle detectors, we should be able to detect the by-products of proton decay.

In essence, this is what several groups around the world have been doing for the past two decades. And the result of all of these years of watching

water? Well, not a single proton decay has been observed. The lower limit of the proton lifetime now appears to be at least 10^{32} years.

Proton Decay Experiments

The IMB experiment — a collaboration between the universities of Irvine, Michigan, and Brookhaven — began in the early 1980s and ran until 1991. The detector was a giant tank of water, lined with sensitive light detectors, based 600 meters underground in the Morton salt mine near Cleveland, Ohio. (See Figure 61.) The setup waited silently for the flash of light that would accompany proton decay. More recent detectors, such as Super-Kamiokande in the Kamioka Mozumi mine in Japan, are even larger. Super-Kamiokande is a cylinder, roughly 40 meters in diameter and 40 meters in height, filled with thousands of tons of extremely pure water. The walls are covered with about 13,000 photomultiplier tubes that are sensitive enough to detect a single photon. (See Figure 62.) Such setups are a more peaceful way of performing experiments than accelerating particles to near light speed and smashing them together!

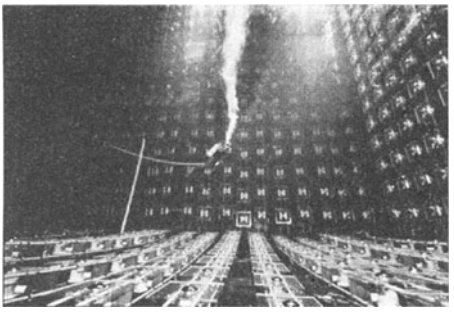


FIGURE 61 A diver inside the water in the IMB proton-decay detector. Although the detector found no evidence of proton decay, it found other interesting physics results.

The failure to observe proton decay is one of those null results in physics that is almost as important as a positive result. A proton lifetime of more than 10^{32} years unambiguously rules out SU(5), or any extensions to SU(5), as a model for grand unification. (Some might also consider this null result to be one of those tragedies of science — a beautiful theory slain by an ugly fact, to use Thomas Huxley's famous phrase. Nevertheless, it is impressive that physicists can test and eliminate grand unification schemes, even though their usual tool — the particle accelerator — can achieve energies nowhere close to the grand unification energy.)

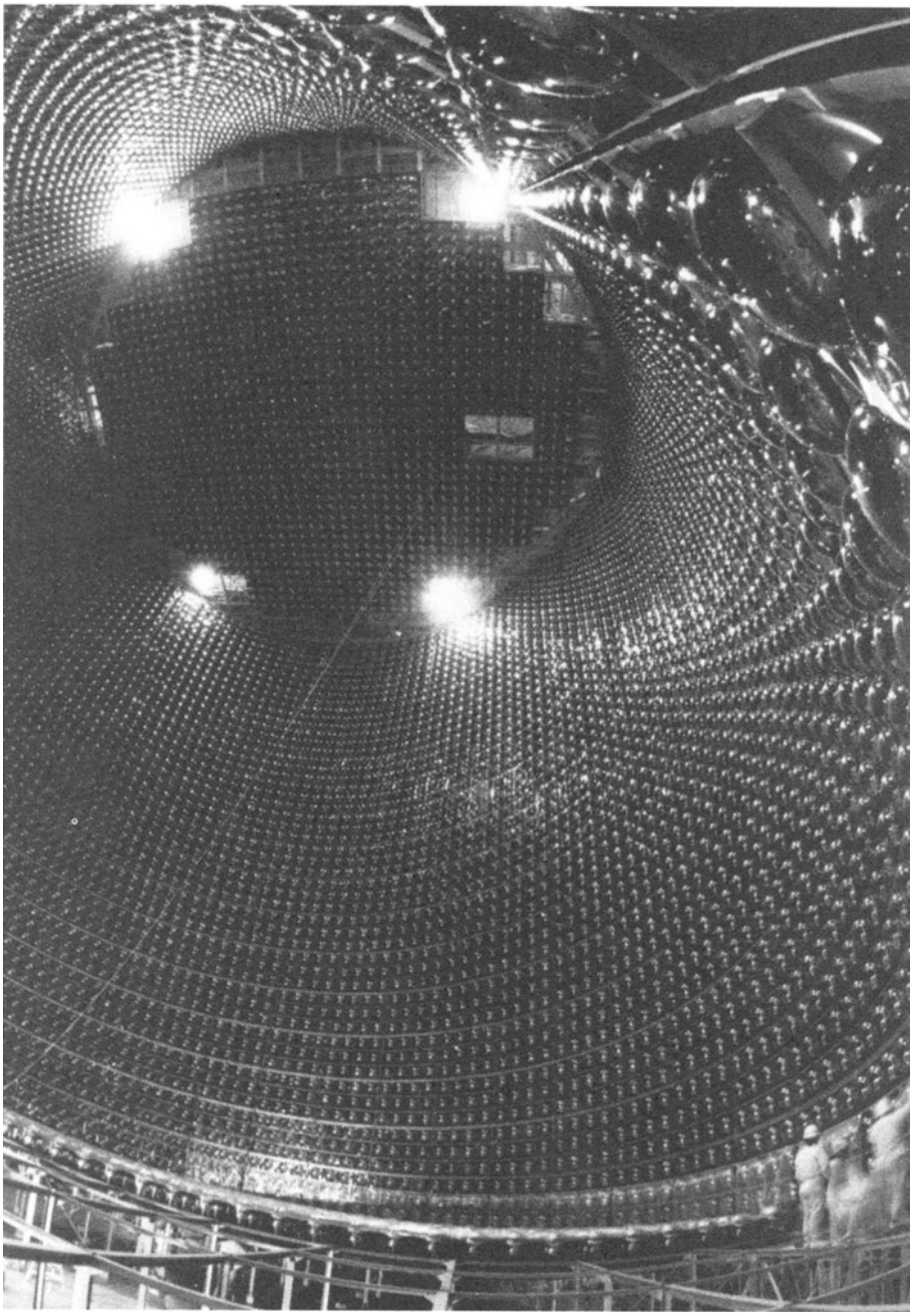


FIGURE 62 A view from inside the Super-Kamiokande detector. The facility was badly damaged in a fire in 2001, but it has recently been rebuilt and is looking again for signs of proton decay. Super-Kamiokande is designed also to look for solar neutrinos (produced by nuclear reactions deep inside the Sun) and atmospheric neutrinos (produced when energetic cosmic rays hit Earth's atmosphere).

Grand Unification and Inflation

The notion of grand unification led to the most profitable idea in modern cosmology: inflation.

In the first 10^{-15} seconds after the Big Bang, the Universe was extremely hot and extremely dense. It was also extremely symmetrical, with the fundamental fields exhibiting the full symmetry of grand unification. As the Universe expanded and cooled, however, the symmetry of grand unification was broken; the Universe underwent a phase transition.

After a phase transition, all the material settles into a new phase. But phase transitions can take time. For example, when pure water is cooled quickly it can remain in the liquid state well below 0 °C. Something similar could have happened in the GUT transition: some regions of the Universe could have undergone a delayed phase transition and remained in the "false" vacuum state for some time. This false vacuum state would have stored vast amounts of energy, which would have been released when the transition to the true vacuum took place.

In 1981 Alan Guth, then a young post-doctoral researcher at Stanford University, used the equations of general relativity to calculate what would happen to the expansion of the Universe if a region got stuck for a short time in such a false vacuum state. The result was dramatic. If the GUT phase transition took place 10^{-35} seconds after the Big Bang, and if it stayed in the false vacuum state for just 10^{-35} seconds before completing the transition, then the region could have expanded by a factor of 10^{25}. Guth called this period of super-rapid expansion "inflation." (Inflation may have been discussed even earlier, in 1979, by the Russian physicist Alexei Starobinsky. But the concept certainly did not become generally popular until Guth published his key paper.) In a modified form, inflation solves many problems inherent in the original formulation of Big Bang theory. Nowadays nearly all cosmological models assume that a period of inflation took place in the early Universe.

The SU(5) model was just the first attempt at grand unification; the surprise is not that SU(5) failed, but that it provided an explanation for so much physics. Although SU(5) is ruled out by experiment, we do not have to give up on the idea of grand unification. There are other symmetries, larger than that described by SU(5), that could provide grand unification. One promising candidate is SO(10), first considered in 1974 by Howard Georgi and independently by Harald Fritzsch and Peter Minkowski, who were at the time both based at Caltech. The SO(10) group contains SU(5)

as a subgroup, but it achieves grand unification in a more elegant and natural manner than SU(5). The SO(10) group is more complicated in that it contains 45 gauge fields, but it provides insight into more parts of physics. It can help explain, for example, why there are three color charges. Another promising candidate for unification, considered first in 1976 by Feza Gursey, Pierre Ramond, and Pierre Sikivie, who were working at Yale University at the time, is based on the exceptional group E(6). Since SO(10) and E(6) both raise the GUT scale and produce proton decay lifetimes that are in excess of 10^{32} years, neither of them have been ruled out by experiment. Grand unification remains on the table.

Grand Unification Based on E(6)

The choice of E(6) as a grand unification symmetry group is an interesting one. As we saw earlier, Lie groups are classified into two types. The first type is the three infinite families SU(N), SO(N), and Sp(N). If N can take *any* number, it is reasonable to ask: why should nature use a particular value from the infinite set of possibilities? Why SU(5)? Why SO(10)? The second type is the five exceptional groups G(2), F(4), E(6), E(7), and E(8). Of these, only E(6) can give rise to a chiral theory (in four dimensions, at least), so it is the only realistic possibility among these groups for a GUT. And it works well. However, it is a large group — it contains SO(10) as a subgroup — and in this model there are 11 additional fermions, including new leptons and quarks. We must assume that, if E(6) grand unification is indeed realized in nature, these new particles are too massive to have been observed.

GUT problems

Putting to one side the non-observation of proton decay, there are other problems with SU(5) grand unification, some of which are generic to grand unified theories.

From the start, many physicists were worried that there appeared to be a desert of sorts between the electroweak scale and the GUT scale. Experimental physicists have a history of finding new particles as they crank up the energy: first atomic nuclei, then protons, then quarks, then W and Z bosons. The prediction of the minimal SU(5) model seemed to be that — with the exception of the Higgs boson — there was no new physics, no new particles, to be found between electroweak energies (of the order of 10^2 GeV) and GUT energies (of the order of 10^{15} GeV). We would have

to increase energies at our colliders by a factor of a thousand billion before seeing anything of interest. If there were a desert in which no new physics appeared and if proton decay were too slow to observe, it would seem to be impossible to put the ideas of grand unification to the test. That would be extremely disappointing. (Mere disappointment is not a valid reason to reject a satisfactory theory, of course. Besides, we need not give up all hope. There was a time when grand unification *must* have played a key role: in the very early Universe, when energies were extremely high. Even if particle accelerators could not traverse the desert, even if proton decay experiments could not find signatures of GUTs, it might still be possible to understand something of grand unification by looking for clues in cosmology. For that reason a close connection has developed between the work of particle physicists and that of cosmologists.)

A more compelling criticism of grand unified theories is that, although *some* of the arbitrary parameters found in the Standard Model can be predicted, by no means all of them can. For example, GUTs give us no idea as to why quarks and leptons have the masses we observe; nor do GUTs explain the number of generations that we see.

More seriously, when we follow carefully the evolution of the three coupling constants of the Standard Model with energy, we see that they do *not* meet at a single point (see Figure 63). They come close, but close is not good enough. Unification does not occur. For a while, this seemed to kill the whole notion of grand unification.

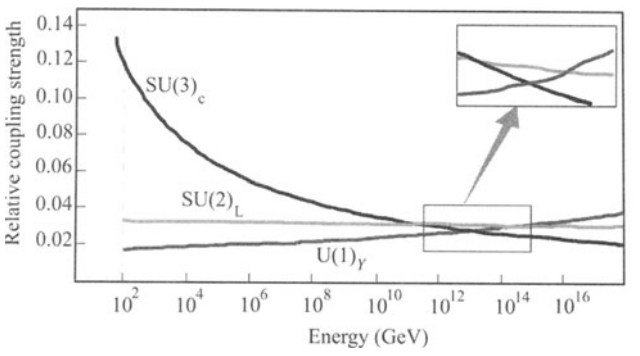


FIGURE 63 A closer examination of the evolution of the three coupling constants in the SU(5) GUT theory shows that they do not in fact meet at a point. They come close, but they miss each other. Grand unification apparently does not take place, at least not in simple theories like SU(5).

Perhaps most seriously of all, GUTs suffer from a terrible affliction known as the *gauge hierarchy problem*. This may seem like a purely technical problem, but attempts to solve it have driven many developments in particle theory over the past two decades. We shall meet it again later in the book; here, let's take a first look at the problem.

In the Standard Model, a fermion develops its mass through its interaction with the scalar (spin-0) Higgs field, which is supposed to pervade

the Universe. The mass of a fermion is proportional to the strength of its coupling to the Higgs field. The Higgs particle itself develops a mass through its self-interactions and its interactions with other particles. However, the fact that the Higgs is a scalar particle causes difficulties. A scalar particle interacts in a completely different way from a particle with non-zero spin. Quantum effects give a scalar particle a mass that is as large as the largest mass scale in the theory, something that does not happen for particles with spin. In a grand unified theory, the largest mass scale is the unification energy, which is about 10^{15} GeV. In other words, according to quantum mechanics, the Higgs particle should have an observed mass of about 10^{15} GeV. However, precision measurements of the electroweak sector at particle colliders clearly indicate that the Higgs particle *must* have a mass that is not too different from the mass of the t quark. More precisely, at the time of writing, the Higgs particle must have a mass greater than 114 GeV and less than 249 GeV. If the mass of the Higgs particle is too large — and 10^{15} GeV is *way* too large — then the usual formulation of the Standard Model itself is wrong.

So, on the one hand, experiment and our understanding of the Standard Model tells us that the Higgs is rather light (of the order of 10^2 GeV). On the other hand, quantum effects tell us that the Higgs must be extremely massive (of the order of 10^{15} GeV). This is the gauge hierarchy problem. Ultimately, it stems from the fact that there are *two* mass scales in GUT models: one associated with electroweak unification and one associated with grand unification. It is difficult to keep these two scales separate for a spin-0 particle when quantum effects play a role.

There is a way around the problem. Before its interactions with other particles, the Higgs particle possesses a "bare" mass. We could fine tune its bare mass so that it cancelled the extra mass due to its interactions, leaving an observed mass of a few hundred GeV. But we would have to fine tune the mass to about the twenty-fourth decimal place in order to maintain consistency with observations. Fine tuning a parameter in order to arrive at a sensible result is an ugly way to proceed; it is more honest simply to hold your hands up and admit that your theory is not working. A more acceptable way around the hierarchy problem is to suppose that, at energy scales around 1 TeV, something new happens — some new physics beyond the Standard Model comes into play. It seems that a desert — which was a feature of early GUT theories like SU(5) — cannot exist.

For all reasonable grand unified theories, the gauge hierarchy problem is a killer. And grand unification still does not incorporate gravity.

But all is not yet lost.

SUPERSYMMETRY

Supersymmetry is the supreme symmetry.
Gates, Grisaru, Roček, and Siegel, *Superspace*

In an earlier chapter we looked at some transformations that can take place in spacetime: *translations*, which shift an object from one point to another; *rotations*, which transform an object by rotating it in space; and *boosts*, which give an object a different velocity. All known laws of physics are invariant under these transformations: we have thus identified some spacetime symmetries. The full set of these symmetries is called Poincaré symmetry, which is named after the great French mathematician Henri Poincaré. (It turns out that, if the symmetry is made local, then a new field, the gravitational field described by general relativity, appears. It is worth bearing this point in mind.)

Under some rather general assumptions, physicists have shown that the Poincaré symmetry is the maximum spacetime symmetry that transforms particles into identical particles. So we would seem to have identified all the possible symmetries to do with space and time. There is, however, a spacetime symmetry that goes beyond Poincaré symmetry. It does so by transforming particles into *different* particles — ones with different spin. This symmetry — *supersymmetry* — transforms fermions into bosons and bosons into fermions. (The spin is always changed by $\frac{1}{2}$; thus a spin-$\frac{1}{2}$ fermion can be transformed into a spin-0 or spin-1 boson.) Other properties possessed by a particle, such as mass and electric charge, are unaffected by the supersymmetry transformation. Under supersymmetry, the clear distinction we have made so far between "particles of matter" (fermions) and "particles of force" (bosons) begins to blur.

Supersymmetry is perhaps the most beautiful of all the symmetries we have looked at so far. And again under some general assumptions, physicists have shown that supersymmetry is the maximum possible extension of Poincaré symmetry. So if we can find evidence for supersymmetry, we will know that nature has utilized all mathematically consistent spacetime symmetries. All this may seem to be of academic interest only, but supersymmetry is a concept that has a whole range of applications.

Theorists stumbled across supersymmetry while they were investigating matters quite unrelated to the grand unification idea we have been discussing. In fact, supersymmetry was formulated independently three times in the early 1970s, by theorists working on both sides of the Iron Curtain — each with no knowledge of the others' work. The Russian physicist Yuri Golfand and his student Evgeny Likhtman discussed supersymmetry in 1971, but their contribution went unnoticed. In 1973, in Ukraine,

Dimitrij Volkov and Vladimir Akulov rediscovered the concept. In 1971, in America, Pierre Ramond and also André Neveu and John Schwarz discovered a form of supersymmetry. (The motivation for their work will become apparent in Chapter 7.) And in 1974, Julius Wess (who was then at Karlsruhe University) and Bruno Zumino (then at CERN) first applied supersymmetry to quantum field theories. (The term "supersymmetry" was coined by Abdus Salam and John Strathdee in 1974.) None of this work was applied immediately to GUT ideas, but it makes sense to introduce the notion at this point because, if nature does indeed possess supersymmetry, then some of the problems with grand unification simply vanish!

Recall that a major failure of grand unification is that the three coupling constants of the Standard Model do not converge on a single value. Another problem is that GUTs suffer from the gauge hierarchy problem.

Well, when one recalculates the evolution of the coupling constants, taking into account the effects of supersymmetry, it turns out that the coupling constants *do* converge on a single value! As shown in Figure 64, a true grand unification is possible in a supersymmetric grand unified theory (SUSY GUT). Furthermore, the contributions to the Higgs mass from interactions with supersymmetric particles exactly cancel the contributions from "normal" particles: with supersymmetry, the hierarchy problem is no longer a problem! All this seems too good to be true, so let's look in a little more detail at what supersymmetry is and what it implies.

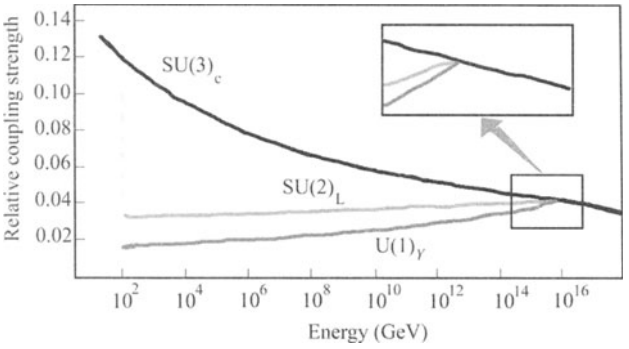


FIGURE 64 When supersymmetry is taken into account, the evolution with energy of the Standard Model coupling constants is slightly altered. This change in behavior means that convergence is much better, and grand unification can indeed take place.

Suppose that we globally transform a fermion into a boson differing by a half unit of spin. Well, if the laws of nature possess supersymmetry then we will find that nature is unchanged by this global transformation. Invariance under transformation: symmetry.

The first and most obvious difficulty with supersymmetry is that none of the known fundamental particles have a supersymmetric partner that allows the transformation to take place. For example, the electron has

completely different properties to the photon, weak bosons, and gluons. It has no superpartner. So how can we possibly talk about supersymmetry as if it were the solution to all our physics problems?

The answer, it must be admitted, is rather weak. We simply *assume* that every Standard Model particle has a supersymmetric partner whose spin differs from the original by a half unit of spin. We further assume that, for some reason, the superpartners have so far eluded discovery. This procedure has every appearance of being a cop-out. (If your outlook on life is of the "glass half full" rather than the "glass half empty" variety, then you could argue that precisely half the number of particles predicted by supersymmetry have been observed: namely, the fundamental particles we have discovered so far.) But we will return to this shortly.

The names of the hypothesized superpartners are derived from the Standard Model particle names by adding the prefix "s" to the fermions and adding the suffix "ino" to the bosons. For example, the supersymmetric partner of the electron is called the selectron; the quarks have squarks as superpartners; the supersymmetric partner of the photon is called the photino; the gluons are partnered with gluinos; and so on. Symbolically, we denote a superparticle by adding a tilde (˜) to the corresponding particle symbol; for example, $\tilde{\nu}$ is the sneutrino, $\tilde{\mathrm{W}}$ is the Wino, and so on. The terminology of supersymmetry is ugly in the extreme, but it seems that we are stuck with it. Table 1 lists the Standard Model particles and their corresponding superpartners.

TABLE 1 The Standard Model particles and their supersymmetric partners.

Particles		Superpartners	
Particle name	Spin	Superparticle name	Spin
electron e	1/2	selectron $\tilde{\mathrm{e}}$	0
muon μ	1/2	smuon $\tilde{\mu}$	0
tau τ	1/2	stau $\tilde{\tau}$	0
neutrino ν	1/2	sneutrino $\tilde{\nu}$	0
quark q	1/2	squark $\tilde{\mathrm{q}}$	0
photon γ	1	photino $\tilde{\gamma}$	1/2
$\mathrm{W}^{\pm}$	1	Wino $\tilde{\mathrm{W}}$	1/2
Z^0	1	Zino $\tilde{\mathrm{Z}}$	1/2
gluon g	1	gluino $\tilde{\mathrm{g}}$	1/2
Higgs h	0	Higgsino $\tilde{\mathrm{h}}$	1/2

Supersymmetry solves the problem that the coupling constants do not converge because, as we go to higher energies, the sparticles begin to have an effect; the presence of the sparticles changes the evolution of the coupling constants. Supersymmetry solves the hierarchy problem because, for every contribution to the Higgs mass from a Standard Model particle, there is an equal and opposite contribution from its supersymmetric partner. (See Figure 65.) The picture of the world described by a supersymmetric grand unified theories is one of remarkable symmetry and simplicity. Think of a quantum particle and imagine that the particle has an arrow associated with it, pointing in some hypothetical space. If the arrow points one way, then the particle is a quark; if it points another way, then the particle is a photon; rotate the arrow again, and the particle is an electron; and so on. Physics does not change as this imaginary arrow turns. There is, in a sense, just one type of fundamental stuff.

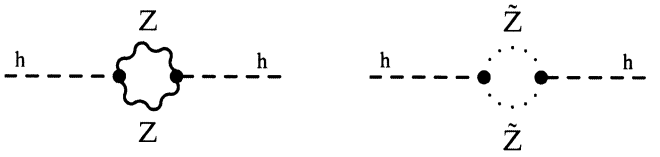


FIGURE 65 The diagram on the left shows a contribution to the Higgs mass from the Z boson. This is exactly cancelled by the diagram on the right, which is a contribution to the Higgs mass from the Zino — the supersymmetric partner of the Z boson. The contributions are equal but opposite. The same mechanism applies to all the particles, thus solving the hierarchy problem.

So SUSY is beautiful, it seems to permit a true grand unification of the three forces, and it solves the gauge hierarchy problem. Anything else? Well, yes, maybe. It might, for example, solve the mystery of the missing matter of the Universe (see the boxed material on page 131). SUSY has a lot going for it

Perhaps the most intriguing aspect of supersymmetry is this. Suppose you repeat a supersymmetry transformation on a particle. For example, suppose you transform a spin-½ fermion into a spin-0 boson and then back into a spin-½ fermion. It turns out that the fermion moves to a different point in space. (See Figure 66.) In other words, repeated application of the supersymmetry transformation gives us a spatial transformation — which is a type of Poincaré transformation. And local Poincaré invariance, as we mentioned earlier, is the symmetry that gives rise to general relativity! We thus see a hint of a link between supersymmetry and gravity, which is remarkable since on the face of it supersymmetry appears to have little to do with gravity.

Supersymmetry and the Missing Matter of the Universe

Astronomers have known for decades that as much as 90 percent of the matter in the Universe is in some dark, as yet undetected, form; it makes its presence felt only through its gravitational influence. This "dark matter" is the gravitational glue that holds together galaxies and galaxy clusters; these structures do not have enough *visible* matter to keep from flying apart. Various lines of evidence suggest that this matter must be in the form of "exotic" particles; in other words, it cannot be made of the protons, neutrons, and electrons that contribute most of the matter around us.

An excellent candidate for the dark matter of the Universe is the LSP — the lightest supersymmetric particle. In most models of supersymmetry, there is a quantity called R-parity that is multiplicatively conserved; all particles are defined to have $R = +1$ and all sparticles are defined to have $R = -1$. What this means is that if, for example, we collide two particles together then we can hope to produce a pair of sparticles; but we cannot produce just one sparticle in such a collision, because that would violate conservation of R-parity. Similarly, if we start with a single sparticle then, if it decays, it must do so into an odd number of sparticles plus an even (perhaps zero) number of ordinary particles; otherwise the conservation of R-parity would be violated. If R-parity is indeed conserved, it follows that the lightest supersymmetric particle must be absolutely stable — because there is no less massive object with $R = -1$ into which it can decay. So any LSPs that were created in the aftermath of the Big Bang must still be around today, contributing to the mass of the Universe but hidden from our view since they do not interact with the photons that we can detect. Which of the supersymmetric particles is the LSP? Well, that is unknown, since we do not know the mass of *any* of the sparticles with certainty. In fact, according to the rules of quantum mechanics, different particles can mix together; in most SUSY models, the LSP is not a pure sparticle at all — rather, it is a spin-½ object called the *neutralino*, which is a quantum combination of the photino, Zino, and Higgsino. The lightest neutralino (symbol $\tilde{\chi}_1^0$) would behave in many ways like a heavy neutrino.

Astrophysics observations may thus provide us with evidence for supersymmetry. Occasionally, neutralinos will collide with other particles, and when that happens an "excited state" of the neutralino can be created. We know that the lightest neutralino, $\tilde{\chi}_1^0$, is stable; but the next-lightest neutralino, $\tilde{\chi}_2^0$, can decay into the $\tilde{\chi}_1^0$ plus an electron–antielectron pair. So the search for high-energy antielectrons in the cosmic background would be one signature of supersymmetry.

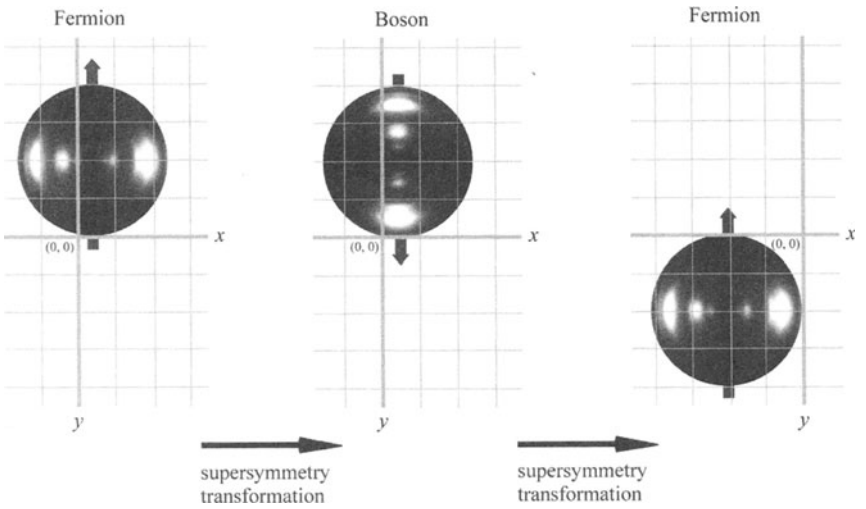


FIGURE 66 The repeated application of a supersymmetry transformation produces a spatial translation! Suppose we begin with a fermion; after a supersymmetry transformation we have a boson, differing from the original fermion by one half unit of spin but at the same coordinates; after a second application of the transformation we again have a fermion with the original amount of spin — but the fermion occupies a different point in space.

We still have not begun to exhaust the theoretical attractions of SUSY. Its attractiveness is one reason why, in the three decades since its introduction, the topic has generated several tens of thousands of research papers — the result of many thousands of person-years of work. Nevertheless, we have to face the unpalatable fact that there is *no* experimental evidence for supersymmetry in fundamental physics. There has been no squeak from a squark; no sign of a selectron. Such a delay in finding experimental evidence for the concept makes many physicists uncomfortable. The experimental verification of general relativity, after all, came just four years after Einstein published his theory; even the elusive neutrino required only 27 years for its detection after Pauli's prediction of its existence. The delay in finding the Higgs boson, which is yet to be observed since its existence was postulated in 1964, is little comfort. That physicists are willing to devote so much time and energy to investigating the properties of a phenomenon that may turn out to be nonexistent is perhaps an indication of how deep, and how beautiful, SUSY is. Regardless of its theoretical elegance, however, most physicists would much prefer to have some direct, tangible evidence for the phenomenon. We have already seen that astrophysical searches for the neutralino might provide evidence for SUSY. Are there any other ways we can search for SUSY?

If we are going to look for supersymmetry we must first accept that it is clearly *not* a symmetry of the world we see around us. Supersymmetry, if it exists, must be badly broken. Spontaneously broken symmetry is a phenomenon we have met before, of course. The electroweak symmetry is not a symmetry of the world we see around us; at high energies the Universe possesses the symmetry, but at everyday energies the symmetry is broken and we perceive distinct electromagnetic and weak forces. Perhaps the same is true of supersymmetry. Perhaps the Universe began in a maximally symmetric state, with the laws of nature being invariant under supersymmetric and gauge transformations. Under exact supersymmetry and gauge symmetry, all matter and gauge boson states would be exactly massless. As the Universe cooled and supersymmetry broke, the sparticles would gain mass, but the ordinary particles of the Standard Model would remain massless due to the unbroken gauge symmetry. Maybe the breaking of supersymmetry then naturally induced the breaking of gauge symmetry. Thus the particles and their superpartners would develop different masses: the sparticles would become heavy while the ordinary particles would remain light. We now see the particles because they possess relatively little mass and are therefore easy to produce; we do not observe the sparticles because they are massive and difficult to produce.

The crucial question is: what are the masses of the supersymmetry particles? If the lightest sparticles have a mass of, say, 10^{15} GeV (typical of the GUT scale), we have essentially no chance of creating them in collider experiments. But if the lightest superpartners to the ordinary particles have a mass of, say, 200 GeV (typical of the electroweak scale), physicists will soon be able to produce copious numbers of them in collisions at the Large Hadron Collider at CERN. When a pair of gluons collide at LHC energies, there is a good chance of producing a pair of gluinos; and when a gluon collides with a quark, there is a good chance of producing a gluino and a squark. Any gluinos and squarks thus produced would decay in unmistakeable cascades to lighter states. We could conceivably even see sparticles before the 2007 switch-on of the LHC in collisions at the Tevatron.

So which is correct: do the lightest sparticles appear at the GUT scale, in which case we will not see them, or do they appear at the electroweak scale, in which case we can hope to see them soon?

At present, no one knows how supersymmetry (if it exists) is broken. Rather than have the symmetry break spontaneously, most models require symmetry-breaking terms to be included explicitly "by hand," which is an unsatisfactory solution. Nevertheless, if we believe that supersymmetry is the solution to the gauge hierarchy problem then it is clear that the sparticle masses cannot be very much larger than the typical particle masses in the

Standard Model. If supersymmetry is exact, then contributions to the Higgs mass from a particle and its corresponding sparticle would exactly cancel; in broken supersymmetry, the contributions do not cancel exactly — but they are of the order of the sparticle masses. We know that the Higgs mass must be of the order of 100 GeV, so the masses of the squarks, sleptons, and the rest cannot be *too* large. If supersymmetry exists, we will likely see it at the LHC within the next few years.

If we do *not* see supersymmetry at the LHC then our theories of physics may need a radical rethink. For the moment, let's assume that supersymmetry will be found and see where that assumption takes us.

Local Supersymmetry

Superspace is the greatest invention since the wheel.
Gates, Kangaroo, Roachcock, and Gall, *Stuperspace*

The supersymmetry we have looked at so far is a global symmetry. If we make a particular transformation at one point, we have to make the *same* transformation at all other points. A global symmetry somehow sounds more important than a local symmetry but, as we have emphasized before, a local symmetry is much more restrictive and therefore more powerful. For local supersymmetry to hold, we require that we can make different supersymmetry transformations at different points and find that everything is unchanged. Quite some requirement!

If the laws of physics are invariant under some global symmetry, it does not necessarily follow that they will be invariant under local symmetry. Promoting a global symmetry to a local symmetry only works if we introduce compensating fields — gauge fields — which then give rise to forces. For example, we learned that the equations of electromagnetism were invariant under a global change of phase; they are also invariant under the much more restrictive *local* change of phase, so long as we introduce a compensating gauge field. This freedom to choose the phase of the Dirac field differently at different spacetime points *requires* the existence of the photon. Under global symmetry we can have electric charges, but only under local symmetry can we have photons and thus light. This is a profound result. You might expect a similarly profound result if you demand that nature should be locally supersymmetric; you might expect to find interesting physics if you ask for the freedom to choose between particle and sparticle differently at different spacetime points. Your expectations would be correct: something profound *does* happen when you demand that local supersymmetry should hold.

Just as in previous cases, you can have local supersymmetry as long as you introduce compensating fields. In fact, if supersymmetry is made local, *two* gauge fields are required. One gauge field communicates information about the supersymmetry transformation; for mathematical reasons this must be a spin-3/2 field. The second gauge field communicates information about the product of two supersymmetry transformations — which, as we saw, was equivalent to a translation in spacetime. It turns out that this must be a massless spin-2 field.

The introduction of a massless spin-2 field sets enough bells ringing to excite a campanologist. If one believes that the gravitational force arises on the quantum scale through the exchange of virtual particles, one can deduce the properties of the exchange particle — the *graviton*. The graviton must be massless because, as with electromagnetism, gravity is a long-range force. Gravity is always attractive, so the graviton must be a boson with an even-integer spin such as 0, 2, 4, ..., since particles with an odd-integer spin (like the photon) produce a repulsive force between like particles. If the graviton were a spin-0 particle, calculations show that we would not observe gravitational lensing — the bending of light rays by mass. But astronomers routinely observe this phenomenon, so spin-0 is ruled out for the graviton. Spins higher than 2 are ruled out for technical reasons to do with the consistency of the theory. We are left with one possibility: the graviton is a massless spin-2 particle. The graviton, we can now see, is one of the gauge particles of local supersymmetry! (The other gauge particle, the spin-3/2 partner of the graviton, is called the *gravitino*.)

It is possible to argue in reverse. If we start from the assumption that the gravitational force arises through the exchange of massless spin-2 particles in a manner described by a quantum field theory, then we can *derive* the equations of general relativity. So merely by demanding that supersymmetry should be a *local* symmetry, we find that gravitational interactions appear automatically. This is another of those wonderful results!

Local supersymmetry is called *supergravity*, since general relativity appears naturally when supersymmetry is made local. The first supergravity theory was developed in 1976, by Daniel Freedman and Peter van Nieuwenhuizen, who were then at the State University of New York at Stony Brook, and Sergio Ferrara, who was then at the Ecole Normale Supérieure in Paris. It was the first important advance on general relativity in the six decades after its formulation. Their theory was the simplest possible supergravity: it contained one graviton, one gravitino, and nothing else. As a model of the Universe, therefore, it was not particularly realistic; we know there are fundamental quark and lepton fields with spin-1/2 and fundamental boson fields with spin-1 and spin-0. Clearly, though, since supersymmetry can

connect any adjacent pair of fermions and bosons, it should be possible to extend the original supergravity theory to include further particles. In fact, there are precisely eight possible *extended supergravity* theories. They are distinguished by the label $\mathcal{N}$, which runs from 1 to 8, with $\mathcal{N}$ being the number of distinct supersymmetry transformations in the theory.

A satisfying feature of supergravity is that the particle content is precisely specified by the restrictions of symmetry. The original theory was $\mathcal{N} = 1$ supergravity, and as we mentioned above it had one graviton and one gravitino; the next most complex case is $\mathcal{N} = 2$ supergravity, which possesses one graviton, two gravitinos, and one spin-1 particle; $\mathcal{N} = 3$ supergravity has one graviton, three gravitinos, three spin-1 particles, and one spin-$\frac{1}{2}$ particle. The theory with the most symmetry of all is $\mathcal{N} = 8$ supergravity. It has a unique gauge symmetry, described by the group O(8). The mathematics of group theory tell us that $\mathcal{N} = 8$ extended supergravity has one graviton, eight gravitinos, 28 spin-1 particles, 56 spin-$\frac{1}{2}$ particles, and 70 spin-0 particles. (There can be no supergravity theories with $\mathcal{N}$ greater than 8 because any such theory would inevitably contain particles with spin greater than 2. Such theories are known to be inconsistent.)

The $\mathcal{N} = 8$ supergravity theory caused great excitement when it was first published. It is not difficult to understand why. The theory seems, at first glance, to possess enough fundamental fields to be a realistic candidate theory of unification. After all, it has 28 gauge bosons — seemingly more than enough force mediators to describe a grand unified theory. Moreover, the theory would not just unify the strong and electroweak interactions; gravity would be included too. And the unification would take place in a completely natural and elegant way, as one might expect of a theory with the maximum possible spacetime symmetry. There was another benefit with supergravity: it could help with the infinities of quantum gravity.

We have already met infinities in quantum theories. For example, we saw earlier how QED, the archetypal quantum field theory, contains infinities. Recall that the lowest-order contribution to a QED process gives a finite and sensible answer. The next-order contribution, which should produce a small perturbation to the original answer, is in fact infinite. The first-order perturbation contains diagrams with one "loop." (The process of an electron emitting and then reabsorbing a photon is an example of a diagram with a loop; see Figure 42.) It is these one-loop diagrams that give an infinite probability. Similarly, the second-order contribution contains two-loop diagrams, and these are infinite too. The third-order contribution contains three-loop diagrams, which are infinite; and so on. In QED, QCD and electroweak theory, these diagrams can be renormalized — swept away into definitions of mass, charge, and coupling constant.

The quantized form of general relativity contains loop diagrams that we cannot remove so simply. There is an underlying difference between quantum electrodynamics and quantum gravitodynamics: in QED the force on a particle is proportional to its electric charge, whereas in gravity the force on a particle is proportional to its mass. But Einstein told us that mass is equivalent to energy, so the strength of the gravitational force increases at high energies. In a quantum process, therefore, the gravitational force increases as the energy of the exchanged virtual graviton increases. It is this high-energy behavior of quantum gravity that causes all the problems. When we calculate the contribution to the probability arising from a loop diagram, we have to sum over all possible energies of the virtual particles. In QED, this is no problem because the force does not depend on energy; in quantum gravity, though, the contribution increases as the energy increases. We can think of the infinities in QED as being of the form $1 + 1 + 1 + \ldots$, which we can handle. The infinities in the quantized form of general relativity are of the form $1 + 2 + 3 + \ldots$, which are much more vicious. There is an infinite variety of infinities, and the theory is non-renormalizable.

Supergravity can help tame the infinities. The original theory of quantum gravity assumed that the gravitational interaction was mediated only through the exchange of gravitons. But in supergravity, the exchange of gravitinos can also contribute to the gravitational interaction. (Gravitinos do not mediate a force by themselves. They are fermions and so obey the exclusion principle; they do not combine to produce a macroscopic force. Gravitinos *can* be exchanged, however; it is just that, being fermions, they must be exchanged in pairs.) At low energies, the probability of a pair of gravitinos being exchanged is negligible; gravitation occurs entirely through the exchange of gravitons, and we have the usual predictions of general relativity. At high energies, however, gravitinos are exchanged in larger numbers; and when we calculate the contributions of diagrams containing gravitinos, we find that the infinities cancel! In fact, all supergravity theories were quickly shown to be finite at the one-loop level. They were soon shown to be finite at the two-loop level, too.

An enticing picture emerges. Perhaps at high energies (much higher than collider energies, but not necessarily as high as the Planck scale) nature possesses the maximum amount of supersymmetry: $\mathcal{N} = 8$ supersymmetry. This could have occurred in a time fractionally after the Big Bang. In $\mathcal{N} = 8$ supergravity we have a unified theory of gravity, particles, and forces. As the Universe expands and cools, supergravity breaks. What is left as a low-energy effective theory would be a supersymmetric grand unified theory. We could have supersymmetry, and thus a solution to the

gauge hierarchy problem (and all those other ills that the supersymmetry medicine cures). Supersymmetry would then somehow be broken, and we would be left with the familiar SU(3) × SU(2) × U(1) of our very-low-energy world. In the early 1980s, several physicists hoped that $\mathcal{N} = 8$ supergravity might be *the* unified theory.

Unfortunately, the picture sketched above does not work. Nature has declined the chance to employ such a beautiful theory — at least in the simple form discussed above. The theory of $\mathcal{N} = 8$ supergravity possesses some fundamental flaws.

FIGURE 67 In 2001, supergravity reached its 25th birthday. The Supergravity at 25 *conference, hosted by the State University of New York at Stony Brook, reunited the three creators of the first supergravity theory — Peter van Nieuwenhuizen, Sergio Ferrara, and Daniel Freedman — along with many younger physicists who continue to investigate the ramifications of the theory.*

First, the gauge group of the theory, O(8), does not contain the Standard Model SU(3) × SU(2) × U(1) as a subgroup. In fact, the *largest* subgroup of O(8) is SU(3) × U(1) × U(1); this is simply not large enough to accommodate the Standard Model. There are ways around this limitation. For example, the theory also has a "hidden" symmetry of SU(8), so it is possible that the SU(2) of the Standard Model is a subgroup of this SU(8). This would imply that the photon and the gluons are truly fundamental, whereas the $W^{\pm}$ and Z^0 are composite. While this is possible, it seems aesthetically unpleasing.

A more serious difficulty is the lack of suitable leptons in the theory. Although $\mathcal{N} = 8$ supergravity contains 56 fermions, after various symmetry breaking operations have taken place there are not enough colorless leptons for the three families that we observe.

Another problem is that all extended supergravity theories are non-chiral. It is extremely difficult to explain in a realistic fashion why we should observe a chiral Standard Model.

The mortal blow to $\mathcal{N} = 8$ supergravity, though, is the fact that it is non-renormalizable. The infinities of quantum gravity are ameliorated by supergravity, but they do not vanish completely. When we calculate graviton–graviton scattering at the seven-loop level, back come those infinities.

We have travelled a long way in this chapter. With the notions of grand unification, supersymmetry, and supergravity, we come tantalizingly close to unifying the microworld with the macroworld. But close is not close enough. If we want to incorporate gravity into our theory of the microworld, it seems that we need new ideas. Interestingly, perhaps the most important of these ideas has been around for almost as long as general relativity itself. We now turn to one of the "wild ideas" of physics: higher dimensions. If we cannot achieve unification in four dimensions, then perhaps we can achieve it in higher dimensions.

6

HIGHER DIMENSIONS

Yet I exist in the hope that these memoirs,
in some manner, I know not how, may find their
way to the minds of humanity in Some Dimensionality
and may stir up a race of rebels who refuse
to be confined to a limited Dimensionality.
E. A. Abbott, *Flatland*

Einstein spent the greater part of his adult life searching for a unified theory — one that would unite gravity with the other forces. At least, that is what he *thought* he was doing. Since the precise nature of the forces acting at the nuclear level was unknown to him, he could only attempt to unify gravity with electromagnetism. We now understand that the weak and strong forces are as fundamental as electromagnetism, and in the six decades since Einstein's death, our knowledge of particle physics has mushroomed. With hindsight we can see that Einstein's quest was doomed. A unified theory of gravity and electromagnetism would, at best, be only part of the story.

Nevertheless, a unification of gravity with electromagnetism would be worth having; ideas that lead to a successful unification of gravity with *any* of the other forces might help us with the problems we encountered in earlier chapters. Remarkably, soon after Einstein published his theory of general relativity, an idea was floated that *did* allow gravity to be combined with electromagnetism. It was a wild idea then, and it remains so today. But if we can accept it, we might just have the key to unlocking the secrets of the Universe.

KALUZA–KLEIN THEORY

To dream of new dimensions.
Robert Graves, *The Castle*

In early 1919, Theodor Kaluza — then a junior academic at the University of Königsberg — wrote to Einstein with a seemingly outrageous proposal. Kaluza wrote down the equations of general relativity in *five* dimensions — four of space and one of time. The extra dimension here simply meant that Kaluza needed five coordinates to specify a spacetime event: four coordinates (instead of the usual three) to specify the spatial position of the event, and the usual time coordinate to specify its temporal location. Such a procedure may sound strange, but there is nothing in Einstein's theory to forbid the introduction of extra spatial dimensions. When Einstein developed his theory, he naturally assumed that three spatial dimensions were all he had to worry about, but his theory makes just as much mathematical sense if we employ extra dimensions. Indeed, this freedom to choose the dimensionality of spacetime might be regarded as a weakness of Einstein's theory: would it not be better if the theory *told* us how many spatial dimensions the Universe possesses? Weakness or no, Kaluza was at liberty to study the equations of general relativity in five dimensions, and when he did so he discovered something very strange. The four-dimensional form of both Maxwell's equations of electromagnetism and Einstein's equations of general relativity appear naturally from the full five-dimensional theory.

Einstein was stunned: "At first glance I like your idea enormously," he wrote to Kaluza. And a little later: "The formal unity of your theory is startling." It was, indeed, a bewildering result. Gravity and electromagnetism appear to have very different properties, yet Kaluza had shown how they both emerge naturally from a single theory. Einstein's heart must have pounded with excitement when he pondered the suggestion. Nevertheless, Einstein delayed publication of Kaluza's paper for more than two years, and it is not difficult to understand why he had misgivings: what on Earth could the extra dimension mean in physical terms?

FIGURE 68 T.F.E. Kaluza

If there is one thing we can be sure about from our everyday experience, it is this: the world possesses three large spatial dimensions. We can move up–down, left–right, forward–backward. That's it. Mathematically, we know that we only ever need *three* coordinates to specify the position of any point in space, just as we need *one* coordinate to specify the position of any point on a line, and *two* coordinates to specify the position of any point on a plane. This implies that our world is three-dimensional. More practically, if we open a bottle of gas and place it in a sealed container, random collisions cause the gas molecules to diffuse into all possible spatial dimensions — and it is clear that they diffuse into only *three* dimensions; we never observe molecules to disappear into some other dimension. So what are we to make of a fourth spatial dimension?

The notion of higher dimensions was "in the air" in the early part of the twentieth century. Mathematicians had realized that they could study the properties of higher-dimensional objects using their usual techniques, and they produced many interesting results. In a sense, when Einstein and Minkowski developed the concept of four-dimensional spacetime, they were introducing these mathematical notions of higher dimensions into physics. So perhaps it was not unreasonable that physicists, seeing the success of the Einstein–Minkowski approach, sought to introduce further dimensions. Indeed, five years before Kaluza's proposal, the Finnish physicist Gunnar Nordström had attempted to unify gravity with electromagnetism by considering a five-dimensional form of Maxwell's theory of electromagnetism. (Nordström abandoned his approach when Einstein published the theory of general relativity. His attempted unification failed because his theory could not explain a key observation: the bending of light in a gravitational field.) So higher dimensions were a hot topic in the physics of the early twentieth century. But was the introduction of a fifth spatial dimension just some mathematical trick, or could it be interpreted physically? The most successful interpretation of Kaluza's idea came in 1926. It was due to Oskar Klein.

Klein realized that Kaluza's five-dimensional theory contained a major flaw: since it was essentially Einstein's theory of general relativity, albeit in a higher-dimensional spacetime, it was a *classical* theory. By 1926, it was clear that the world obeyed the laws of quantum mechanics. So Klein decided to determine whether Kaluza's theory was compatible with quantum mechanics. If it were not, the theory would have to be abandoned.

Klein began by writing the fundamental equations of quantum mechanics in five dimensions — four of space and one of time. He solved the equations, showing how one could view the solutions as being electromagnetic and gravitational waves in our usual four-dimensional spacetime. The idea of a fifth spatial dimension thus passed the test that Klein had posed: the idea was compatible with quantum mechanics. But the introduction of quantum mechanics suggested a physical interpretation of the fifth dimension.

Suppose, Klein said, that the fifth dimension exists as a real, physical dimension — and that it is circular. In other words, suppose there is a circle at each and every point in spacetime. Most people (I am certainly one of them) find it difficult, if not impossible, to visualize such a thing. A simple analogy, however, may help to clarify our mental picture. Instead of the full three dimensions of our world, consider a mathematically idealized line — something that possesses only one dimension, which means that we need just a single number, or coordinate, to identify any point on the line. Now suppose that, to each and every point on the line, we associate a circle. In order to locate a position on the circle we need another number, or coordinate. So we have a two-dimensional object. This may sound rather complicated, until we realize that we have simply given an idealized description of a garden hose. A circle at every point on a line generates a cylindrical object, like a hose. It is more difficult to imagine associating a circle to every point in our Universe, but the principle is the same. According to Klein, our Universe has a five-dimensional structure. We need four coordinates to specify a point in our familiar spacetime (three space coordinates and one of time, as specified by Einstein) and at each point we need a coordinate to specify a position on the associated circle. See Figure 69. (The garden-hose analogy is not perfect. The hose exists *in* a four-dimensional spacetime. The five-dimensional spacetime described by Klein need not be "in" some yet higher-dimensional space.)

Viewed from a distance, a garden hose looks like a line. If we noticed an ant crawling on the hose, we would say that the ant could move in only one dimension: either forward or backward along the hose. It is only as we get closer to the hose that we see it has a second, circular dimension. It is only close up that we see that an ant, in addition to moving *along* the hose, can also move *around* the hose and return to its starting point. Klein's suggestion is that a similar picture applies to our Universe. The reason we do not see the fifth, circular dimension is that the radius of the circle is small. If the radius of the circle were much less than the size of a gas molecule, for example, then it would explain why gas molecules do not diffuse into the extra dimension — they are simply too big to do so. According to the uncertainty principle we can investigate small distances

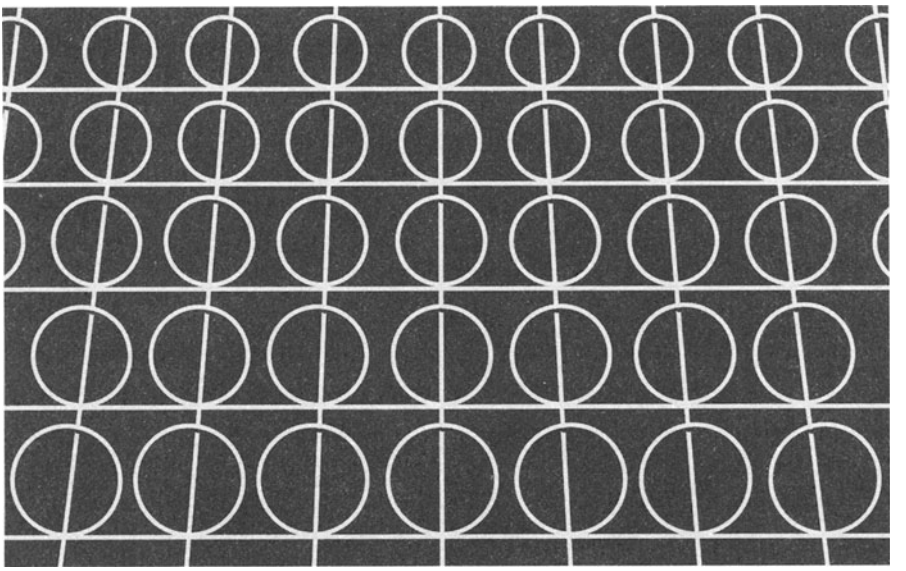


FIGURE 69 Oskar Klein suggested that there is a tiny circle at every point of our Universe. So spacetime has five *dimensions, rather than the four we are familiar with (three space dimensions, plus time). The reason we do not see the extra dimension is that it is too small to probe; it is far smaller than protons, for example, so even subnuclear particles can behave as if the fifth dimension does not exist.*

only by using high-energy particles as probes. If the energy of a particle is small then its associated wavelength is large — and thus it may wash over the circular fifth dimension. At our highest-energy accelerators we can probe distances of roughly 10^{-18} meters. So if the diameter of the fifth dimension were much smaller than 10^{-18} meters, we would not notice it even in our most sophisticated experiments.

Because of Klein's improvement on Kaluza's idea, the generic name of Kaluza–Klein theory is given to this method of unification. The original Kaluza–Klein theory gives rise to a beautiful view of the Universe. The underlying five-dimensional spacetime possesses only *one* force — the gravitational force. Everything is simple. But observers like us at large distances, in a four-dimensional spacetime, see something quite different. We see gravitation, which is the curvature of four-dimensional spacetime. And we see electromagnetism, which is actually the vibrations of the underlying gravitational field in the circular fifth dimension. (We should also see a scalar field, and its associated spin-0 particle called the *dilaton*. In Kaluza–Klein theory, the dilaton is as natural an outcome as the photon.)

Supermarkets often tout "buy one, get one free" offers. Well, if you are willing to buy the idea of a compact, circular fifth dimension, then you

get electromagnetism for free — Maxwell's theory of electromagnetism is an inevitable consequence of Einstein's geometrical theory of gravity. (As a special offer, you get the dilaton field too, but it is not clear what value to place upon that.) As we shall soon see, with fairly straightforward extensions to the idea we can incorporate the other gauge forces — the weak and strong forces — into a single unified picture.

FIGURE 70 Oskar Klein

Different people have different ideas of what constitutes a bargain. For some, the idea of introducing an unseen fifth dimension is too high a price to pay simply in order to solve a problem in physics. For my own part, I find the ideas associated with extra dimensions inherently less bizarre than, say, the ideas of quantum mechanics. The difference between these two sets of ideas, of course, is the quality of the supporting evidence. Every experiment we perform tells us that the Universe is quantum mechanical in nature; we *have* to accept the weirdness inherent in quantum mechanics, whether we like it or not. In contrast, there is *no* experimental evidence that might lead us to believe in the reality of higher dimensions. Nevertheless, the equations of general relativity make sense with a fifth dimension; Klein's idea — that the fifth dimension is compact — explains why we do not see it. There is also an argument coming from cosmology that at least hints at the possibility of extra dimensions (see the boxed text on page 147). So there is no immediate reason for rejecting the notion of higher dimensions.

All this is mere handwaving, but it makes it reasonable for us to at least believe in the *possibility* of the existence of higher dimensions. If we do this, we find a strong sense of freedom: we are free to write down our theories not just with one extra dimension, but with *any* number of extra dimensions. We can then investigate mathematically which theories are consistent, which theories make internal sense. If none of the theories are consistent, we could reject the whole notion of extra dimensions. But maybe some of the theories *do* make sense and *do* deserve further theoretical study. Of course, we would be honor-bound to make contact with the real world eventually: any such theory would have to reduce to the Standard Model at low energies, explain why we do not see the extra dimensions, and make experimental predictions that we could someday hope to test. In the short term, though, it is not unreasonable to take the Kaluza–Klein idea seriously and see what we can make of it.

A Cosmological Argument for Extra Dimensions

The Universe is expanding. In the very early stages of the Universe, the large flat dimensions that we now see must have been very small and highly curved. So suppose that the Universe began with five compact dimensions, each of the order of the Planck size of 10^{-33} meters. We know of no reason why the initial expansion would have to take place at the same rate in each of the dimensions. We can imagine that one of the dimensions stayed small, while the other dimensions expanded. We could end up with spatial dimensions of the order of 10^{26} meters (roughly the present radius of our Universe) and an orphan dimension of 10^{-33} meters. The same argument applies to any number of dimensions; the Universe may have started with *many* small dimensions — it's just that only a few of them expanded into the large Universe we see today.

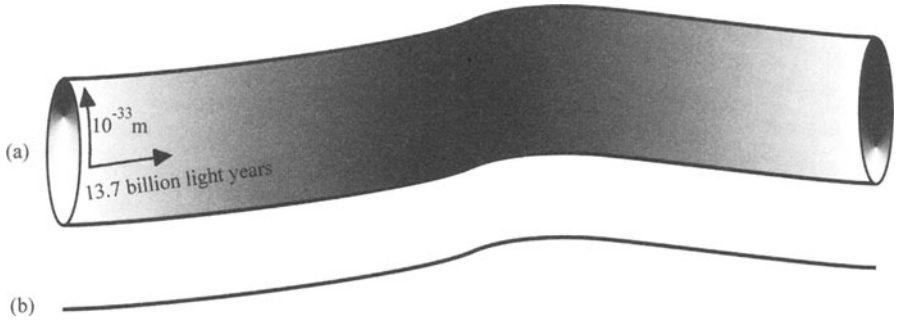


FIGURE 71 Suppose that a spacetime started out with two spatial dimensions, each with a size of 10^{-33} meters (two Planck-size dimensions, in other words). And suppose that after a big-bang event, billions of years ago, one of the dimensions started to expand — and continues to expand — while the other remains small. We could picture the spacetime as looking something like the figure in part (a). For an observer in this spacetime who can access only low energies, the spacetime would look like the figure in part (b); the extra dimension would be effectively invisible.

The Kaluza–Klein idea is now an old one; it predates gauge field theories, grand unification, supersymmetry, and several other topics we have discussed. But after an initial surge of interest upon publication of Klein's paper, physicists dropped the topic for several decades. If it is such a good idea for unification, why did physicists not embrace it more readily? Perhaps Kaluza–Klein theory has a fatal flaw?

We have already dealt with the first objection to the Kaluza–Klein idea, namely: why do we not see the fifth dimension? We do not see it because it is compact. The natural question to ask is then: exactly how large is

the extra dimension? Klein himself first came up with an argument to estimate the size of the compact dimension — but the result just highlighted a practical problem for Kaluza–Klein theory.

Klein's argument is roughly as follows. According to quantum mechanics, the waves associated with particles can move around the circular fifth dimension. In order to be observed in experiments, however, the waves must be stationary, in the same way that the waves on a guitar string must be stationary because the string is fixed at both ends. So the wave either has a constant amplitude, or else it has a wavelength equal to the circumference of the circle (which means the wave has precisely one oscillation around the circle), or half the circumference (two oscillations), or a third of the circumference (three oscillations), and so on. The energy, and thus the mass, of the particles is inversely proportional to the wavelength. So the lowest-mass particles in Kaluza–Klein theory are the ones with infinite wavelength — those with constant amplitude, in other words. These particles have zero mass. The first particles with mass in Kaluza–Klein theory are those with a wavelength equal to the circumference of the circle; precisely one oscillation fits around the circle.

We can estimate the mass of these particles as follows. Since the theory unifies gravity and electromagnetism, the particle has an electric charge as well as a mass; in fact, the electric charge is inversely proportional to the circumference of the circle. If we assume that this charge is equal to the charge on the electron, then we can calculate the mass. It turns out to be about 10^{16} times the mass of the proton. This is far too large for us to observe in our collider experiments. (In Kaluza–Klein theory there is an infinite tower of particles, with each particle corresponding to an integer number of oscillations around the circle. Finding evidence for this tower of particles would bolster the theory; but how can we find evidence for the tower of massive particles if we cannot produce even the least massive of their number?) If the first massive particle is indeed 10^{16} times the proton mass, we can estimate the circumference of the circle: it is about 10^{-32} meters, which is not too far from the Planck length. And this is the problem: such a small system seems to be impossible to investigate in practice. Certainly, it is not surprising that physicists in the 1930s were unable to think of direct experimental tests to check the idea. In contrast, the ongoing developments in quantum field theory offered lots of scope for experimental verification. Physicists were naturally attracted to the ideas of quantum field theory and repelled from the notion of extra dimensions.

The dilaton, the scalar field that comes with the theory, proved to be another obstacle to the acceptance of the Kaluza–Klein idea. Physicists nowadays are more sanguine about postulating the existence of hitherto unobserved fundamental particles — it worked spectacularly for Pauli with

the neutrino, after all — but back in Klein's day such behavior was altogether more frowned upon. Suppose we take the dilaton seriously, though, as some brave physicists did. The problem is that the dilaton would be a massless particle, or at least extremely light. Presumably, such particles would have been created in large numbers in the immediate aftermath of the Big Bang, and we would expect to observe them. We can easily enough account for the non-observation of sparticles, X bosons, and the rest of the menagerie of hypothetical particles: we simply say that their large mass puts them beyond the reach of our present-day detectors. But to explain the non-observation of a low-mass particle like the dilaton is more tricky. This was not a fatal problem for the theory, but it was embarrassing.

There was a worse problem with the dilaton. The size of the extra dimension at any point depends upon the dilaton field at that point. In Klein's initial formulation he took the dilaton field to be a constant; in other words, the radius of the circle is the same at each point. But there is no good reason why it should have a fixed size from the outset. It is more natural to let the size of the extra dimension vary from spacetime point to spacetime point. If we do that, the dilaton field does not take on a nice fixed value; it tends to go either to zero or to infinity. In turn, this means that the extra dimension either "pinches off" to zero at some points or "blows up" to infinity at other points. Imagine standing on a rubber hose that is carrying a flow of water; the "extra dimension" of the hose gets pinched off at the point where your foot stands, while somewhere else in the hose the water pressure causes the "extra dimension" to balloon. Clearly, nowhere in the Universe do we see the tiny extra dimension of Klein start to balloon and grow large. It is perhaps less clear that we do not see the dilaton field shrink to zero size — but theorists have shown that if the field *were* zero then we *would* notice it.

For all these reasons, and more, physicists lost interest in Kaluza–Klein theory. The idea of extra dimensions is, however, bewitching; the unification of gravity with electromagnetism seems almost magical. In the late 1970s, the feeling came about that maybe Kaluza and Klein — and Nordström before them — had come up with an idea before its time.

Yet More Dimensions...

One's mind, once stretched by a new idea, never regains its original dimensions.
Oliver Wendell Holmes

In the original Kaluza–Klein theory, the U(1) local gauge invariance of electromagnetism comes about because of the circular geometry of the extra dimension. Relativity is invariant under rotations around a circle — and

the group that describes invariance under rotations around a circle is precisely U(1). Suppose that, instead of a tiny circle at each point, there were *two* tiny circles; this would mean that spacetime had six dimensions. In this case, the symmetry group would be U(1) × U(1), and an observer in our large four-dimensional world would observe *two* different types of photon — one for each invariance under separate rotations around the two circles. We can generalize this idea. Instead of there being a single tiny circle at each spacetime point, suppose that at each point there is a compact space with many dimensions. In the jargon, suppose there is a *manifold* at each point. This may seem absurd — and certainly it is impossible to depict accurately on the page — but is it any more absurd to postulate the existence of, say, a compact seven-dimensional space at each point than it is to postulate the existence of a compact one-dimensional space at each point? If we are willing to accept that our Universe has many higher dimensions, then we obtain a huge payoff: certain special symmetries of the manifold will appear as gauge symmetries in the large four-dimensional space we inhabit! So if there were a manifold with SU(3) × SU(2) × U(1) as a symmetry group, then we would observe SU(3) × SU(2) × U(1) gauge symmetries — in other words, the Standard Model — in the real world. We would have unified, in a beautiful geometric way, gravity with *all* the other forces. We would have realized Einstein's dream of a unified theory.

In 1981, Edward Witten, a mathematical physicist at the Institute for Advanced Study in Princeton, published a classic paper (one of many).

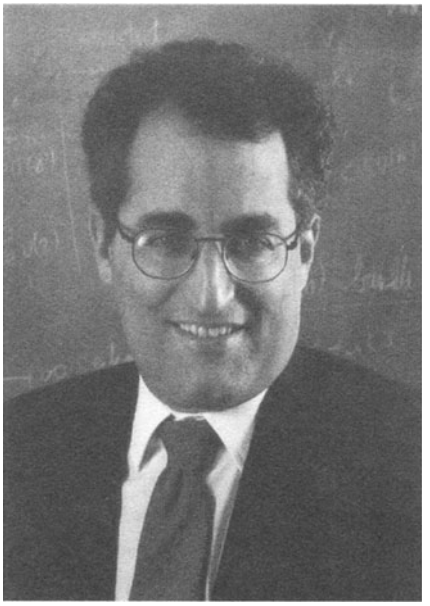


FIGURE 72 Edward Witten is widely regarded as the most brilliant of present-day theoretical physicists. For more than two decades he has set the agenda for investigations of high-energy physics. His work is highly mathematical — perhaps too mathematical for him ever to win the Nobel Prize. He has, however, been awarded the Fields Medal, which is the equivalent for mathematicians of the Nobel Prize.

In it he showed that the *minimum* number of dimensions that the compact manifold could possess, and still have SU(3) × SU(2) × U(1) as a symmetry, is seven. (The number seven comes about because U(1) is the symmetry group of the circle, which as we know has dimension one. A two-dimensional sphere is the space with the lowest dimension that has symmetry SU(2). And it turns out that the lowest-dimensional space with symmetry group SU(3) has four dimensions. The minimum number of dimensions for the compact space is thus 1 + 2 + 4 = 7.) So if our "real" world is an 11-dimensional spacetime, with a compact seven-dimensional space at each point of our familiar four-dimensional spacetime, then we have the *hope* of finding a realistic theory that unifies gravity with the other forces in the spirit of the Kaluza–Klein idea.

As Witten pointed out in his paper, the *minimum* number for the dimensionality of spacetime, 11, is remarkable. It is remarkable because the *maximum* number of dimensions in which one can write down a consistent supergravity theory is also 11. (If you have more than 11 dimensions it turns out that, when you reduce to our familiar four-dimensional Universe, you must observe particles with spin greater than two. But as we saw in the previous chapter, it is impossible to construct a consistent quantum field theory containing gravity if such high-spin particles are allowed.) So if you want your higher-dimensional theory both to contain the Standard Model and at the same time be a supergravity theory then it *must* be an 11-dimensional theory!

Even before Witten wrote his paper, Eugène Cremmer, Bernard Julia, and Joël Scherk, who were all based at the Ecole Normale Supérieure in Paris, had constructed a wonderful theory: $\mathcal{N} = 1$ supergravity theory in 11 dimensions. Their theory was unique. It was the most symmetric, the most beautiful of all supergravity theories. It reduced *automatically* to four large dimensions and a compact seven-dimensional space, thus explaining why we appear to live in a four-dimensional Universe. And upon reduction, the four-dimensional spacetime contained $\mathcal{N} = 8$ supergravity, which, as we saw in the previous chapter, was considered to be a good candidate for unification of the forces in four dimensions. (In fact, Cremmer, Julia, and Scherk were able to write down the $\mathcal{N} = 8$ supergravity theory only by writing down the $\mathcal{N} = 1$ supergravity theory in 11 dimensions and then reducing to four; the higher-dimensional theory is much easier and simpler to work with.)

It seems inconceivable that nature would pass up the chance to make use of this structure. Indeed, for a while, many physicists believed that $\mathcal{N} = 1$ supergravity in 11 dimensions was the unified theory they had been working towards for so long.

Unfortunately, it does not quite work.

FIGURE 73 Eugène Cremmer. Along with his colleagues Bernard Julia and Joël Scherk, Cremmer developed $\mathcal{N} = 1$ supergravity theory in 11 dimensions. For a while, this theory was a candidate "final theory."

In the previous chapter, we saw the problems with $\mathcal{N} = 8$ supergravity. For example, although it is much better behaved than other quantum theories of gravity, it is not renormalizable: when Feynman diagrams with more than seven loops are considered, the infinities do not cancel. (This is perhaps an indication that the theory can work at low energies, but not at high energies; it thus might be an effective theory, but it cannot be a truly fundamental theory.) Witten, in his 1981 paper, pointed out another problem. Because the extra dimensions of space are symmetric with respect to left-hand and right-hand spinning particles, the particle reactions that a Kaluza–Klein-type theory describes do not distinguish between left-handed and right-handed particles. Yet the particle reactions we actually observe are asymmetric when it comes to spin: the electroweak interaction treats left-handed and right-handed particles differently. This is the *chirality problem*. It seems to be merely a technical difficulty, but it is extremely hard to circumvent it. With his paper, Witten had raised great hopes only to dash them a few pages later.

With $\mathcal{N} = 1$ supergravity in 11 dimensions, seven of which dimensions compactify in the spirit of Kaluza and Klein so that they are unobservable, we are *tantalizingly* close to a truly unified theory. We *almost* have a realistic framework for treating gravity on the same footing as the other forces. But we are not quite there. This is too good a theory to discard completely, though. In particular, the notion of higher dimensions should surely be retained. To make further progress, however, we clearly need some fresh ideas.

7

SUPERSTRINGS

Take but degree away, untune that string,
And, hark! what discord follows.
William Shakespeare, *Troilus and Cressida*

We should pause a moment to take stock of what we have learned. In the Standard Model we have a framework, based upon the beautiful principles of quantum field theory and local gauge symmetry, that successfully explains all the particle physics experiments that have ever been performed. But the Standard Model is incomplete: it contains two different sectors that seem to demand unification, and it completely neglects gravity. The first problem is addressed by grand unification, which expands the gauge symmetry of the Standard Model. Grand unified theories themselves suffer from problems, but the problems become much less severe when we introduce another type of symmetry: supersymmetry. And if we make supersymmetry a local symmetry, then we find that gravity appears automatically. We are thus led to supergravity theories. Finally, if we introduce the notion of higher spatial dimensions — an old idea for physicists who have tried to describe gravity and electromagnetism in a unified framework — and combine it with supergravity, we are led to a theory that is *almost* a unified theory of all the fundamental particles and their interactions: 11-dimensional supergravity.

Unfortunately, the theory of 11-dimensional supergravity does not quite work. Its successes perhaps indicate that this is an *effective* theory that works at some particular energy, in much the same way that the Standard Model is an effective theory that works at the energies that are presently available to us. But it cannot be the final theory we have been looking for. Ultimately, the problem with this whole approach seems to be that it is based on quantum field theory, which deals with point-like objects. This approach works spectacularly for quarks and leptons, but it fails when the graviton — the exchange particle of gravity — is involved. When physicists calculate Feynman diagrams that contain gravitons, they eventually encounter infinities. The presence of infinities in a theory is a clear sign that something is wrong. And the infinities cannot be renormalized away as they can for quantum electrodynamics (QED) and quantum chromodynamics (QCD). A quantum field theory that includes gravity yields nonsensical answers.

Eventually, physicists realized that there was another approach to the problem of unifying gravity with the other forces in a quantum mechanical framework. When physicists found this approach — and so far it is the *only* approach that seems to work — many of them were willing to accept its high price. The price is that the familiar point-like objects of quantum field theory — quarks, leptons, gauge bosons, and so on — must be replaced by one-dimensional strings. Furthermore, we have to assume that the strings exist in a higher-dimensional spacetime. And supersymmetry is a prerequisite. If we refuse to accept strings, higher dimensions, or supersymmetry, then there simply does not seem to be enough "room" to accommodate both gravity and the other interactions in a unified quantum theory. Perhaps one day someone with more imagination than physicists of the previous century may discover a way of unifying the various interactions in a way that does not require strings or higher dimensions. But for the past twenty years this idea — strings in higher dimensions — has looked the best bet for unification.

(It should be noted that there is at least one other promising area of research into the problem of reconciling general relativity with the principles of quantum mechanics. It is called *loop quantum gravity*; see the boxed text on page 155 for more details. This approach has had some successes but, since it does not attempt to incorporate the Standard Model of particle physics, it is a less ambitious project than string theory. Since string theory is the only theory that attempts to unify *all* fundamental physics — gravity, quantum mechanics, the Standard Model — we will not mention loop quantum gravity any further. The reader who is interested in learning more about the loop quantum gravity approach is referred to the relevant items in the bibliography.)

This chapter describes the string idea. We start with the unusual history of the string concept and end with a discussion of why string theory is *still* not the final theory physicists hoped for.

Loop Quantum Gravity

Loop quantum gravity (LQG), has been developed since 1986 by several theorists including Abhay Ashtekar of Pennsylvania State University, Carlo Rovelli of the University of Pittsburgh, and Lee Smolin of the Perimeter Institute in Waterloo, Canada. The LQG approach is fundamentally different from the string approach to developing a quantum theory of gravity.

In most versions of string theory, calculations take place in a fixed "background space." Such a space is rather like a stage upon which matter and energy move like actors. But this is a rather Newtonian view; Einstein's general relativity tells us that there is no such thing as a fixed background space. The approach taken by loop quantum gravity theorists, therefore, is to construct a quantum field theory of gravity in a form that does not require a background space.

In LQG, there is no space as we commonly understand it; reality instead consists of "loops" — quantum excitations of the lines of force of the gravitational field. These loops combine and interact to form so-called "spin networks," and these networks evolve into the smooth space with which we are familiar.

The picture emerging from LQG is very different from that coming from string theory, and it is not as well developed. (This may be due as much to sociological difficulties as problems with the theory itself. String theory rose out of studies of high-energy physics, which historically attracted large numbers of researchers and grants; LQG grew out of studies of general relativity, which historically has attracted fewer researchers and grants.) Nevertheless, there is now a thriving community of LQG theorists. If supersymmetry is not observed, if the proton refuses to decay, and if higher spatial dimensions are not found, then surely many more theorists will start to investigate loop quantum gravity.

As with string theory, there are experimental tests of the theory that one might hope may soon be within reach. For example, space is quantized in loop quantum gravity: it has a granular structure at the Planck scale. This could show up in violations of Lorentz symmetry at very high energies — photons with different wavelengths would travel at different speeds, an effect that is forbidden in special relativity. Such effects are currently being sought (see page 274).

THE EARLY HISTORY OF SUPERSTRINGS

A series of failures may culminate in the best possible result.
Gisela Richter

The 1960s were a frustrating decade for physicists studying the strong force. Their experiments had produced hundreds of hadrons — the strongly interacting particles — and each of the hadrons had slightly different properties. The frustration arose because physicists were unable to identify the fundamental hadrons, which in turn meant that they could not develop a quantum field theory of the strong interaction. We now know, of course, that *none* of these hadrons were fundamental. It was only when physicists identified colored quarks as the basic building blocks of hadrons that they could construct a successful quantum field theory of the strong interaction. In the absence of a good theoretical understanding of the strong interaction in terms of fundamental fields, some physicists sought to understand the strong interaction using different approaches than quantum field theory. In 1968, for example, Gabriele Veneziano, a young Italian-born physicist based at CERN, wrote down an equation that described several of the properties of hadrons as observed in scattering experiments. Two years later, several physicists working independently — Yoichiro Nambu at Chicago University, Tetsuo Goto at Nihon University in Tokyo, Holger Nielsen at the Niels Bohr Institute in Copenhagen, and Leonard Susskind, then at Yeshiva University in New York — interpreted Veneziano's equation in physical terms. They argued that the equation described the behavior of one-dimensional objects — "strings." The picture coming from this approach was that hadrons were strings rather than point particles. When hadrons interacted, they did so via the joining and splitting of strings rather than via the exchange of other point particles. This picture had a certain elegance. Rather than there being a plethora of different particles, there was just one basic type of "stuff:" the string, different modes of vibration of which gave rise to the observed particles.

Logically, it seemed there could be two types of string. First, a string could be *open* — a line segment with two free ends. Charges such as electric charge and color would reside at the two endpoints of the string. Nowadays, we call a theory containing open strings a type I string theory. Second, a string could be *closed* — a tiny loop, with no free ends. In this case, charges would be smeared over the whole of the string. We now call this a type II string theory. (We can imagine the two endpoints of a single open string joining to form a single closed string, so a type I theory inevitably contains both open and closed strings. But a closed string can never break to form an open string, so a type II theory contains *only* closed strings.)

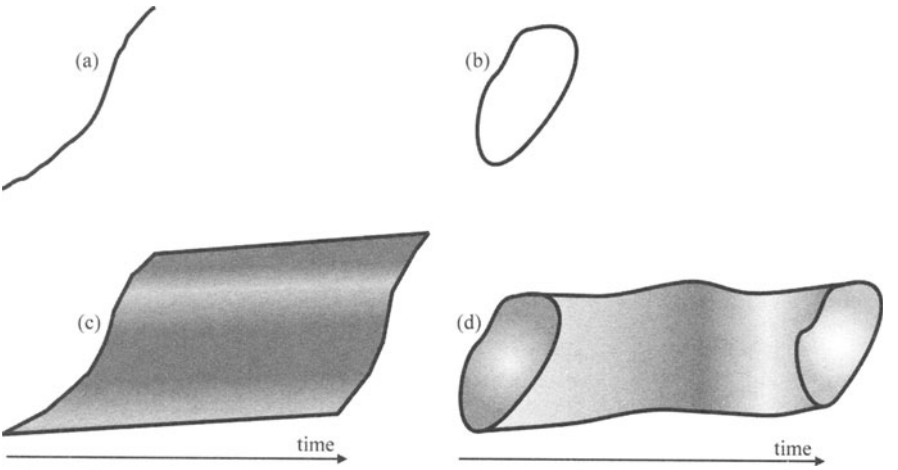


FIGURE 74 A string can be open, as in (a), in which case it has two free ends. Or a string can be closed, as in (b), in which case it forms a tiny loop with no free ends. As a string moves through spacetime it sweeps out a world sheet, a two-dimensional surface that is analogous to the one-dimensional world line swept out by a particle moving through spacetime. Part (c) illustrates the world sheet swept out by an open string; (d) illustrates the world sheet swept out by a closed string.

Figure 74 illustrates the two types of string. Parts (c) and (d) of the figure represent the motion of a string through spacetime. As a string moves through spacetime it sweeps out a two-dimensional *world sheet*, just as a point particle traces out a one-dimensional *world line*. The world sheet of an open string is rather like the surface of a piece of paper; the world sheet of a closed string is like the surface of a cylinder.

A string, as envisaged in this early work, would have a stiffness or *tension*, just as a guitar string or a piano string has a tension (though, unlike musical instruments, there would be nothing anchoring the ends of the string — tension would be an intrinsic property of the string). If a string were "plucked," it could vibrate in various harmonics. The fundamental mode of vibration would correspond to one type of hadron; the first harmonic would correspond to a different hadron; the second harmonic would be yet a different hadron; and so on. The energy scale of the string would depend upon its tension. For example, a piano string is stiffer than a guitar string and it requires correspondingly more energy to set it vibrating. So the fundamental energy scale for hadrons would be determined by the string tension, and higher harmonics of the vibrating string would correspond to more massive hadrons.

It is important to appreciate that the Veneziano-inspired string was an extended object, an object with size. The string size — which would have to be about 10^{-15} meters, the typical scale for hadronic physics — would

be an *intrinsic* property of a string. Compare this with, say, the proton, which has a size because it consists of a spatially distributed set of point-like constituents. At particle accelerators we can "look inside" a proton and see its constituent quarks. But a string would be quite different. A string would not consist of anything smaller or more basic. It would be fundamental, and it would make no sense to ask what the string was made of. All the arrows of explanation, which we mentioned in the Introduction, would converge on the string and then stop.

FIGURE 75 Yoichiro Nambu (shown here on the right, in conversation with Yoshihide Sakai) helped develop the idea of spontaneous symmetry breaking in particle physics and contributed to the development of the quark model of hadrons. He was also one of the first to investigate relativistic string models.

Strings were an interesting idea. Until this idea surfaced physicists, with very few exceptions, had considered fundamental objects only in the framework of *point* particles. Yet as we have seen, points — objects of zero size — cause problems. The infinities that plague quantum field theory can often be traced to the point-like nature of elementary particles; indeed, if an object has zero size, then infinities will inevitably occur. This is illustrated in Figure 76, which shows the most primitive Feynman diagram: two particles coming together and emitting (or absorbing) a third particle. The diagram possesses a discontinuity. By patching together two or more of these primitive vertexes, we can generate diagrams that represent more complicated processes. The problem is that, because the vertex occurs at a point, all observers — even those in relative motion, like observers A and B in Figure 76 — agree precisely when the interaction occurred. It turns out that, because of this property, a certain class of Feynman diagram gives an infinite answer. For the field theories in the Standard Model, we can renormalize away the infinities and the final result is acceptable. But the renormalization procedure does not work for gravity. If one of the particles involved in Figure 76 is a graviton, new infinities appear at each order of perturbation. To repeat: the ultimate source of the difficulty is the point-like nature of the interaction.

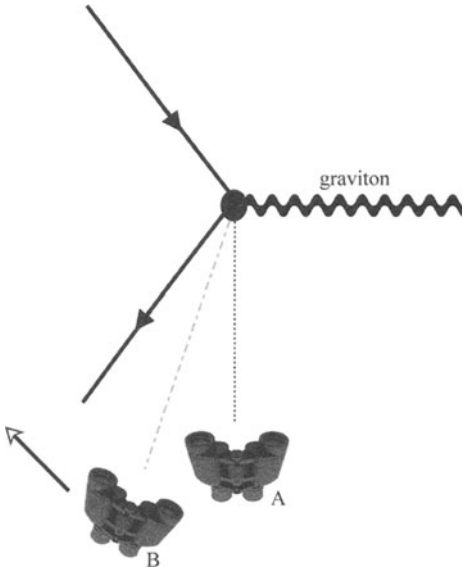


FIGURE 76 This shows a field theory vertex involving a spin-2 graviton. By patching together two or more of these primitive vertexes, we can draw diagrams that represent physical processes involving gravity. Unfortunately, we soon find ourselves drawing diagrams which give rise to infinities that cannot be renormalized away. The theory gives nonsensical answers. Ultimately, the problem arises because of a discontinuity: the interaction takes place at a point. From observer A's point of view, observer B is in motion. But both observers agree about precisely when the interaction took place. The point-like behavior of the spin-2 graviton in field theory is what causes all the problems.

Now consider what happens when two strings join to form a third string, as in Figure 77. There is no discontinuity in the diagram; everything happens smoothly. In this case, one cannot define the precise instant at which the two strings join because the definition of the time coordinate of

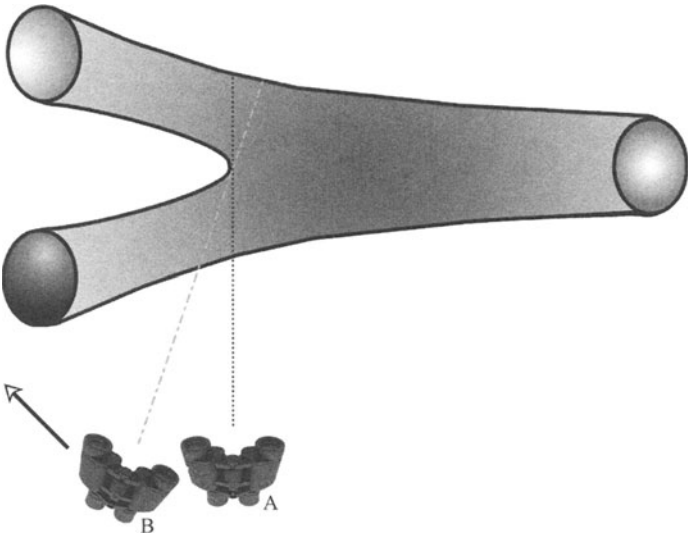


FIGURE 77 This diagram shows the interaction of two strings — the string theory equivalent of a field theory vertex. In this case observers who are in relative motion do not agree when the two strings first interact. There is no unambiguously defined moment when the two strings interact. The interaction is "smoothed out." There is no discontinuity, and the infinities that plague field theory do not arise.

a point on the world sheet depends on the observer. In other words, two different observers can disagree about the instant at which two strings join. This "smooths" the theory, and the infinities do not occur.

In order to compute the probability of any particular string process occurring, one presumably would have to use a perturbative approach, just as we have to do with Feynman diagrams in the Standard Model. In string theory, higher-order diagrams correspond to the number of holes (or handles) in the world sheet. For example, Figure 78 shows a typical string process to three orders in perturbation theory.

FIGURE 78 A typical process in string perturbation theory. At each order, a handle appears. There is only one diagram at each order in perturbation theory, but it is exceedingly difficult to calculate diagrams with more than two handles.

One of the nice features of Figure 78 is that, at each order in perturbation theory, we have only one diagram to contend with. Compare that with the point particle diagrams of the Standard Model, where the number of different diagrams mushrooms at higher orders. It is not all good news, though: the mathematics for dealing with diagrams containing more than about two handles is horrendously difficult. Nevertheless, for weak coupling, the perturbation approach gives us a means of understanding some of the properties of string theory. (The nonperturbative aspects of the theory would be hidden from this approach — just as the property of quark confinement, for instance, is beyond the reach of perturbative QCD.)

Nambu, Goto, Nielsen, and Susskind considered the interaction of relativistic strings. In their model, experiments taking place at accelerators such as those at CERN involved not the collision of point particles but instead involved the collision of strings, which could interact and create new strings. Their string model seemed to provide a reasonable description of hadronic interactions. In addition to this encouraging result, the high-energy behavior of their model was better behaved than its competitors. At the energies at which particle accelerators operated, the strings would appear point-like: the traditional viewpoint and the stringy viewpoint would yield the same results. But at much higher energies, the two pictures diverge: from the traditional viewpoint, higher and higher energies probe smaller and smaller distances, thus introducing an infinity; from the string-based viewpoint, higher energies simply excite the string into different modes of vibration. The high-energy regime of the theory was altered, and the infinities could be avoided.

The string model seemed as if it could become a promising theory of hadrons. Of course, there were problems.

One of the problems was that the original string model described only bosons. The bosonic string could describe the mesons — particles like the pion and the rho — but it simply did not apply to fermions. The bosonic string thus failed to acknowlege the existence of important particles like the proton and the neutron. In 1971, in order to circumvent this problem, André Neveu and John Schwarz, who were then at Princeton University, and, independently, Pierre Ramond, developed a variant of the theory — *spinning string* theory — that could accommodate fermions.

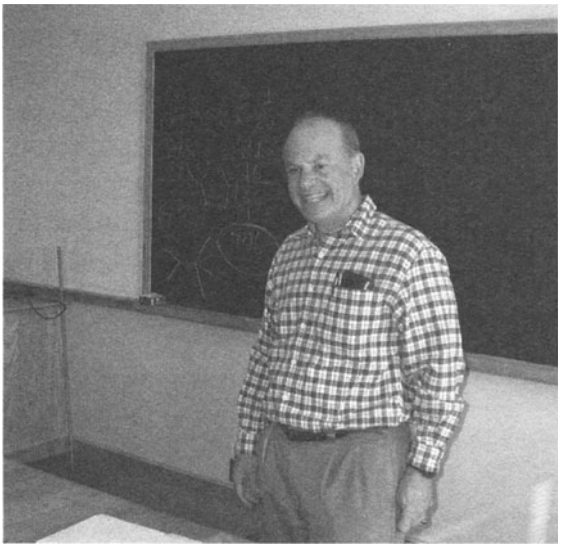


FIGURE 79 John Schwarz is a professor of theoretical physics at Caltech. He has been one of the leading developers of string theory for 35 years.

A second problem was that string theory only made sense if spacetime had many higher dimensions. The reason for this, as with so many things in physics, has to do with symmetry — in this case a beautiful symmetry called *conformal symmetry*. As mentioned earlier, when a point particle moves through spacetime it traces out a world line. An important principle, which is closely related to Noether's theorem and which gives us an elegant way to describe the trajectory of particles, is that a particle moves along the world line that minimizes its *action*. (Action is a number with the dimensions of energy multiplied by time.) The action of a particle is directly proportional to the length of its world line.

We can treat the motion of a string in a similar fashion. Classically, a string moves through spacetime in such a way as to minimize its action. The action of a string is directly proportional to the area of the world

sheet swept out by the string, so if the action has the minimum allowed value it follows that the surface area of the world sheet must also have the minimum allowed value. Now, the area of a world sheet cannot depend upon the way we choose to label the coordinates on the sheet; we must be able to transform the coordinates however we wish and get the same answer for the area. This symmetry — the invariance of physics under transformations of world sheet coordinates — is the conformal symmetry mentioned above.

Unfortunately, when we create a quantum theory based on these ideas, we find that certain processes break the conformal symmetry. Such a process is called a *conformal anomaly*. The presence of an anomaly — any kind of anomaly — is a sure sign that a quantum theory is broken; the theory is inconsistent, and it simply fails to make sense. However, one sometimes finds circumstances in which all the anomalies cancel. If the anomalies cancel, the quantum theory *is* consistent and you can at least progress further with the ideas. Intriguingly, all the conformal anomalies of bosonic string theory cancel — but only if the string lives in a spacetime of precisely 26 dimensions. Similarly, all the anomalies of the spinning string theory cancel — but only if the string lives in a spacetime of precisely 10 dimensions. There is no "wiggle room" in these results. For the theories to make sense, spacetime has to possess either 26 dimensions (for the bosonic string) or 10 dimensions (for the spinning string), and the strings must vibrate in all the spatial dimensions. In a sense, this is an exciting result. For the first time, we have a mathematical consistency condition that *determines* the number of spacetime dimensions. Neither of these numbers is 4, which is the number of spacetime dimensions we observe, but this is not necessarily a drawback in principle: we have seen how, decades earlier, Kaluza–Klein theory postulated the existence of higher spatial dimensions and explained their non-observation. Nevertheless, demanding a spacetime of 10 or 26 dimensions seemed rather excessive for a theory of hadronic interactions.

A third problem with string theory was that it seemed the lowest-energy states of the string were *tachyons* — particles that travel faster than light. This was considered a sure sign that the theory was unrealistic. Yet another problem was that quantum mechanical effects can cause energy cancellations in vibrating strings. The lowest-mass particle in string theory would, naively, have an energy proportional to the string tension; but in some cases the energy cancellation is perfect and what remains is a massless particle. Now a massless particle is fine if you are interested in the photon, perhaps, but it is a nuisance when you are developing a theory of the hadrons, all of which possess mass.

Finally, by 1973, there was another problem for string theorists: the strength of the competition. This was the killer blow. In 1973, quantum

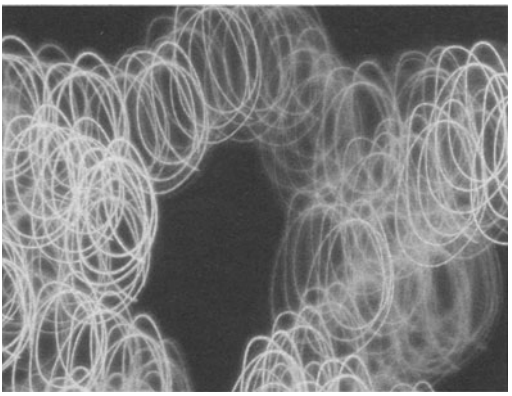


FIGURE 80 In string theory, the fundamental objects are one-dimensional strings; these strings interact with one another, and their vibrations give rise to the familiar particles of the standard model. However, if string theory includes fermions as well as bosons, then the strings must exist in a ten-dimensional spacetime. So any attempt to visualize strings, such as this image, can at best give only a hint of the underlying reality.

chromodynamics was recognized to be the correct theory of the strong interaction, at least at the energies available in colliders. Physicists could understand hadrons simply in terms of bound states of point-like quarks. QCD quickly succeeded in describing hadronic physics in ways that string theory could not match, so it is not surprising that most physicists dropped string theory and instead got into QCD. As fewer minds wrestled with the intricacies of strings, work on the idea sputtered out.

Two physicists who continued to investigate the string idea were John Schwarz and Joël Scherk. They were intrigued by the beautiful symmetries displayed by consistent string theories; they were convinced that such rich mathematical structures must have *some* use in physics. In 1974, they showed that among the massless string states is one with spin 2. Far from being a hindrance to string theory, they showed that this massless state interacted in the same way as the graviton. (This discovery was made independently, and at about the same time, by Tamiaki Yoneya at the University of Tokyo.) So string theory seemed to include general relativity; indeed, it seemed to *demand* the existence of gravity! And other forces could be identified as other massless string states.

It was possible to estimate the string tension from the graviton mode, and it turned out to be the huge figure of 10^{39} tons. (Compare this with the average tension on a steel piano string of 160 pounds; the overall tension on a piano, when it is tuned to concert pitch, is only about 21 tons.) Such a tremendous tension would cause the string to contract to the Planck length of 10^{-35} meters. Schwarz and Scherk made a bold suggestion: maybe string theory had been applied to completely the wrong problem. In a bid to explain hadron physics, everyone had assumed that the typical string size was 10^{-15} meters; instead, if the typical size of a string was of the order of the Planck length, 10^{-35} meters, then string theory had the capacity to

explain *all* the fundamental interactions! It was a candidate for a unified theory!

At first sight there seems to be an obvious problem with this idea. The fundamental mode of vibration of the string would give rise to a particle with a mass of the order of the Planck mass of 10^{19} GeV; the next mode of vibration would give rise to a particle with twice the Planck mass; the next mode to a particle with three times the Planck mass and so on. But the string vibrations are supposed to give rise to particles like the neutrino and the electron, which as we know have small masses — extremely small masses compared to the Planck mass. There is a possible answer: the same quantum mechanical cancellations that give rise to states with zero mass, such as the graviton, can also give rise to states, like the electron, which have a non-zero (but nevertheless tiny) mass. In the case of the graviton, the cancellation is perfect; in the case of the electron, the cancellation would have to be only nearly perfect. If string theory really were a unified theory then it would have to explain not only the mass of the electron, but also the strange pattern of masses of all the other fundamental particles in the Standard Model. Schwarz and Scherk did not know how to do this, and to this day no one knows how to do this. So this answer — that the masses of the Standard Model particles arise from imperfect cancellations in string theory — has not been shown to be correct; but perhaps the details of this question are something we can leave to future generations. At least we can imagine the *possibility* that vibrating strings can give rise to all the particles we think of as being fundamental. Vibrating strings could give rise to all the particles of the Standard Model — the quarks, leptons, and gauge bosons — and in the same framework they would give rise to gravity. It was an exciting proposal.

The physics community, however, was less than overwhelmed by the unification proposal made by Schwarz and Scherk. A few theorists did continue to investigate the properties of string theory and demonstrated a great deal of courage in doing so; it is difficult to pursue an idea when everyone else is telling you the idea leads nowhere. And a few advances *were* made. In 1976, for example, Scherk in collaboration with Ferdinando Gliozzi, then based at the Ecole Normale Supérieure in Paris, and David Olive, then based at CERN, suggested that the spinning string state could be made fully supersymmetric — that there could be equal numbers of bosons and fermions. They also showed that the new theory was free from the tachyons that plagued the bosonic string. But after this, research in strings was almost abandoned. (Scherk made further contributions to physics; as we saw in the previous chapter, he was one of the authors of $\mathcal{N} = 1$ supergravity theory in 11 dimensions. However, he died in 1980 due to tragic circumstances at the age of 30.)

FIGURE 81 John Schwarz (seated far left) and Michael Green (in the foreground, seated next to him) maintained an interest in string theory during several years when most activity in theoretical physics was elsewhere. In 1984, they published a paper that led to an explosion of interest in superstring theory.

John Schwarz, the main defender of string theory, found another collaborator: Michael Green, an English physicist who was then based at Queen Mary College, University of London. Green and Schwarz continued their lonely work on the investigation of string theory. Along with Swedish physicist Lars Brink, they further developed the idea of the spinning string theory with supersymmetry and invented *superstring theory*. There were, in fact, different superstring theories, depending on the amount of supersymmetry possessed by the theory. With $\mathcal{N} = 1$ supersymmetry, the theory could be identified as a type I theory; with $\mathcal{N} = 2$ supersymmetry the theory could be identified as a type II theory. Like its parent, superstring theory made sense only if it lived in ten dimensions. (And, like its parent, it was essentially ignored by most of the physics community, although some people took note. In 1982, for example, Witten read a review article on strings written by Schwarz. Witten classes the realization that the existence of gravity is implied merely by requiring superstring theory to be self-consistent as the "greatest intellectual thrill" of his life. For the most part, though, string theory was ignored.) Then, in August 1984, Green and Schwarz hit gold. They published a paper that led to the "First Superstring Revolution."

THE FIRST SUPERSTRING REVOLUTION

The most heroic word in all languages is revolution.
Eugene Debs

After chipping away at the problems surrounding superstring theory and producing a structure that seemed to possess many advantages for the unification of gravity with the other interactions, string theorists confronted a problem that seemed to kill superstrings stone dead. The problem was that

any superstring theory had to obey some empirical constraints. An obvious constraint, for example, was that there had to be only four *observable* dimensions. No problem there: compactification à la Kaluza–Klein was a way of meeting that constraint. Another, much tighter, constraint was that the theory had to lead to the observed chirality of the weak interaction.

As we saw earlier, the weak interaction has a "handedness," or chirality. We observe certain effects that are a consequence of the weak force, but we never observe the mirror-reflected counterparts of those effects. It is rather tricky, in a four-dimensional spacetime, to construct a quantum gauge theory that is chiral. When you write down such a theory, you can generally find Feynman diagrams that cause the gauge symmetry to break. This in turn leads to unwanted violations of some of our most cherished laws of physics — processes that violate conservation of electric charge, for example. If this happens the theory has a *chiral anomaly*, and you may as well consign the theory to the garbage. It turns out that there are anomalous diagrams in the Standard Model, and these indeed have the potential to ruin the theory. Fortunately, every anomalous diagram in the Standard Model is cancelled by a different diagram. In fact, this is why quark generations must come in pairs, like u and d, or s and c: anomalous diagrams containing the strange quark, say, are always cancelled by similar diagrams containing the charm quark. The Standard Model is thus free from chiral anomalies; it is consistent.

It may be difficult to construct a chiral theory in four dimensions, but it is nigh on impossible to ensure that higher-dimensional theories are both chiral and yet free from chiral anomalies. In fact, most physicists thought it was indeed impossible to construct an anomaly-free superstring theory in ten dimensions; if they were correct, chiral anomalies would kill the whole idea of superstrings.

Now, no one disagreed that it was at least possible in principle to write down a chiral theory in a spacetime of ten dimensions. If fact, it is only possible to define the notion of chirality when spacetime has an even number of dimensions (or, more precisely, when space has an odd number of dimensions; we assume that there is only one time dimension). Again, this is because of symmetry. Chirality depends on what happens after successive mirror reflections along all the spatial axes. If space has an odd number of dimensions then there are an odd number of mirror reflections to consider, and a left-handed object is transformed into a right-handed object. This is obvious when there is only one dimension, and with a little manipulation you can convince yourself it is true in our three-dimensional world. If space has an even number of dimensions, then successive mirror reflections leaves the handedness of any object unchanged. So superstring theory, which lives in a space of nine dimensions (and thus a spacetime

of ten dimensions) at least has the possibility of being chiral. Unfortunately, even if a higher-dimensional theory starts off chiral, when the extra dimensions are compactified chirality is generally lost. You can only maintain chirality in the four spacetime dimensions of our world if the initial higher-dimensional chiral theory *already* includes gauge symmetry groups in addition to the gravitational force. (This is unlike the pure Kaluza–Klein case, where the symmetries we observe come from the symmetries of the compact space.)

What Green and Schwarz showed, in 1984, was that it was possible to formulate a particular superstring theory in ten dimensions that was both chiral and yet free from chiral anomalies. But the requirements of symmetry placed fierce demands on the theory, and in nearly all cases chiral anomalies destroyed it. In fact, superstring theory could only exist if the gauge symmetry group was SO(32). Green and Schwarz also showed that the anomalies would cancel if the gauge symmetry group was E(8) × E(8), although they did not have a specific superstring theory in mind that could contain such symmetry. So only two groups are allowed for consistent superstring theories: SO(32) and E(8) × E(8). Their achievement may seem narrowly technical, of interest perhaps to a handful of people, and yet it created an avalanche of interest in superstring theory — the so-called First Superstring Revolution. To appreciate why their result was so important, why it sparked a resurgence in superstring research, let's consider again what has been achieved.

Imagine that we were to start out with a very simple idea: that the fundamental objects in the Universe are one-dimensional strings, which interact quantum mechanically by splitting and joining together. Merely by attempting to set up a consistent theory based on this idea, a theory that is not immediately incompatible with what we observe, we would be led to a theory with an extremely rich set of features.

We would have a theory that inevitably contains supersymmetry — one of the most beautiful symmetries in particle physics.

The theory would be tightly constrained. It would have no adjustable external dimensionless constants, and it would have just one fundamental energy scale — the string tension. So there would be little room to tweak or fiddle with the theory.

The theory would seem to be immune from the infinities that plague quantum field theories of point particles; we get a perturbation theory that is finite, order by order.

The theory *demands* the existence of gravity, so we would have the prospect of unifying gravity with the other forces — a problem physicists had struggled with for decades.

The theory would be more elegant than the point-particle viewpoint: rather than having a plethora of different fundamental "stuff," we would have only *one* fundamental entity, the string, with the different modes of its vibration giving rise to the different particles we observe. (Since a string can vibrate in an infinite number of modes, there must be an infinite number of "elementary particles." This is not the problem it might at first sight seem: the higher modes of vibration will correspond to particles with a mass of the order of the Planck mass — they will thus be unobservable. The presence in the theory of these ultra-massive particles, however, helps tame the infinities of field theories.)

We would have a prediction for the dimensionality of spacetime: ten. We would thus be led to the idea of higher dimensions, six of which must be small enough to be unobservable.

The gauge group in the full ten-dimensional theory can provide all the observed symmetries of our four-dimensional world. Simply by requiring that the theory is free from anomalies, we would have a tight constraint on the unifying gauge symmetry group. From the *infinite* number of possible symmetry structures, only two are allowed. The unifying group is not *quite* unique — it can be either SO(32) or E(8) × E(8) — but it comes close.

Finally, suppose we were to take the low-energy — or large-distance — limit of string theory. (This might still be a high energy compared to what we study in particle colliders, but it is a low energy compared to the Planck-energy vibrational modes of the string.) At large distances we would not see the extended, one-dimensional nature of the string; we could therefore reasonably approximate the one-dimensional string with a zero-dimensional point particle. In this approximation we would need to use the quantum field theory of point particles, which we looked at in earlier chapters. In other words, the true high-energy theory would be superstring theory; but the low-energy *effective* theory would be a point-particle quantum field theory. And the effective low-energy theory of superstring theory is — supergravity in ten dimensions! Supergravity cannot be the final unified theory because it is non-renormalizable; but it might be the effective low-energy theory of the true unified theory, which is superstring theory. In this picture it seems that the supergravity pioneers of the 1970s were getting the first glimpses of a unified theory: superstring theory. (There is a niggling question here: by far the best formulation of supergravity occurs in 11 dimensions, but such a dimensionality is ruled out by superstring theory. In the superstring approach it would seem that this unique, rather wonderful theory is merely a mathematical curiosity, with no role to play in physics. In the euphoria of the First Superstring Revolution, however, this was a difficulty that most physicists could live with.)

To many physicists involved in the First Superstring Revolution it seemed that nature might have picked out superstring theory as perhaps the only theory that could unify gravity with the other forces. For a while — and, it must be said, not for the first time — the general theoretical physics community felt that it might have its fingertips on a unified theory.

The Fruits of the Revolution

There is geometry in the humming of the string.
Pythagoras

Physicists involved in the development of the Standard Model varied in their opinions about the superstring revolution. Glashow, who first proposed electroweak unification, was scathing. Superstring theory, he said, was "an interesting sociological example of the tendency for physicists to jump on a theoretical bandwagon." On the other hand Salam and Weinberg, who completed the unification first sketched by Glashow, were more enthusiastic. Indeed, after they heard of the discovery of anomaly cancellation, they immediately began working on superstring theory. They were in the majority. Physicists from around the world descended on superstring theory like locusts on a field of corn.

Soon after the groundbreaking work by Green and Schwarz, a group of theoreticians from Princeton University — David Gross, Jeffrey Harvey, Emil Martinec, and Ryan Rohm, who became collectively known as the "Princeton String Quartet," — had constructed a new type of superstring. They called it the *heterotic string*. The name came from the word *heterosis*, a term used in genetics to mean the tendency of cross-breeding to produce an animal or plant with a greater hardiness than either of the parents. The name was appropriate because the heterotic string was a hybrid mix of the bosonic string and superstring, and possessed the strongest features of both. The heterotic string was a weird concoction, but it held out the hope of being *the* unified theory that physicists were looking for.

The Princeton Quartet based their ideas on a string that was closed (so charges were smeared out over the whole string, rather than being isolated as at the two ends of an open string) and *orientable* (which means you can tell in which direction you are travelling as you move along the string). Vibrations can travel around the string in two different directions: clockwise and anticlockwise, or equivalently in right-moving modes and left-moving modes. The key idea is that we treat the two modes of vibration of a heterotic string in quite different ways; the right-movers and left-movers inhabit different worlds. In the heterotic string, the right-movers are super-

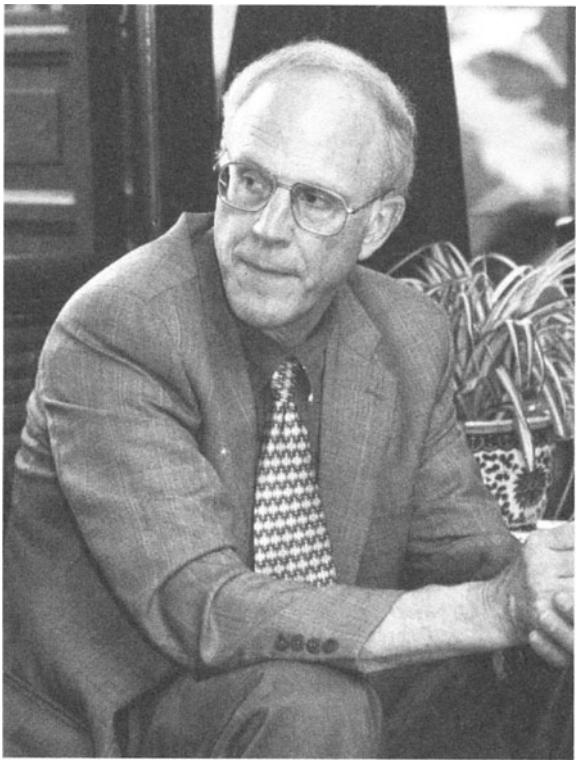


FIGURE 82 David Gross — along with Frank Wilczek and, independently, David Politzer — discovered that QCD possesses the property of asymptotic freedom. Following this discovery, QCD gained widespread acceptance as the correct theory of the strong interaction. Later, Gross was the most senior of the "Princeton String Quartet" — the group of theoreticians that developed heterotic string theory. For a while, physicists clung to the hope that the heterotic string might provide a unified theory of gravity and the other fundamental interactions.

symmetric and live in a 10-dimensional spacetime; they are like the closed superstrings considered so far. The left-movers are non-supersymmetric and live in a 26-dimensional spacetime, like the original bosonic string. To get the heterotic string to make sense, however, the 26 dimensions of the bosonic string must reduce to a 10-dimensional spacetime. The way the Princeton String Quartet did this was to think of the 26-dimensional spacetime as the product of a 10-dimensional spacetime and a "curled up" 16-dimensional space. The final structure is then an object that lives in a 10-dimensional spacetime, like the superstrings we have considered so far. However, reasonable restrictions on the way that 16 dimensions curl up places strong demands of the theory; if the theory is to be consistent, it has to possess one of two symmetry groups: SO(32) or E(8) $\times$ E(8). These are precisely the groups singled out by Green and Schwarz!

There are thus two heterotic string theories living in ten dimensions: the SO(32) heterotic string and the E(8) $\times$ E(8) heterotic string. Now, ten dimensions is six more than we observe. So to make contact with realistic physics, six of the dimensions must be compactified to a small enough size that we do not observe them. The many symmetries of superstring theory

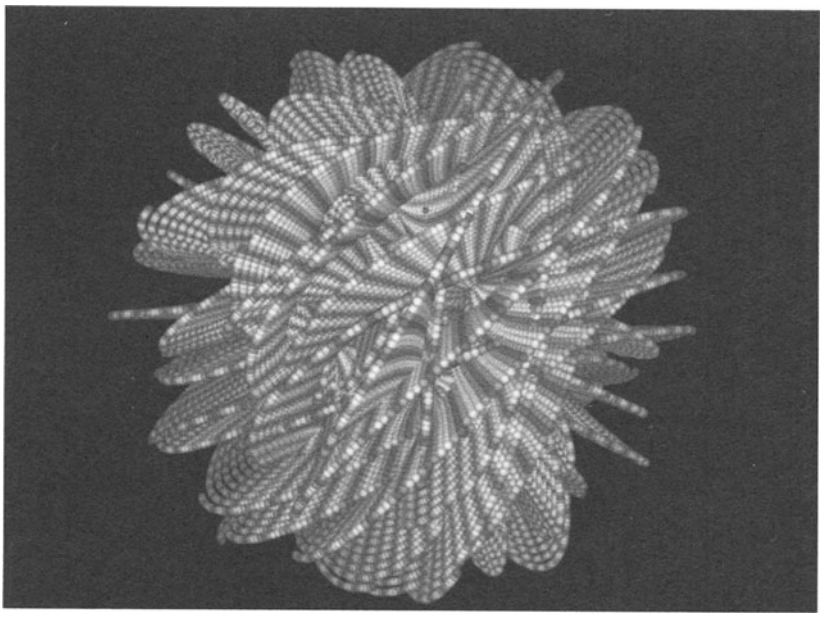


FIGURE 83 An example of a six-dimensional Calabi–Yau space. This is a two-dimensional image that attempts to give a three-dimensional perspective of a higher-dimensional object! Inevitably it is possible to give only a slight flavor of the complexity of a Calabi–Yau space.

demand that the compactified space must possess certain very restrictive properties. In fact, there are good reasons to believe that the space must be of a type called a Calabi–Yau manifold (or else a generalization of such a manifold, called an orbifold). The mathematical existence of this special type of space was first conjectured by Eugenio Calabi in 1954, long before string theory was contemplated. In 1976, Shing-Tung Yau proved Calabi's conjecture and thus showed that these spaces existed. The original Kaluza–Klein theory asked us to picture a small circle at each and every spacetime point. We have to do something similar in superstring theory, but instead of a circle we have to imagine a Calabi–Yau manifold at each spacetime point.

When the extra dimensions are compactified, gravity would still exist in our low-energy four-dimensional world; compactification cannot destroy the gravitational interaction. Furthermore, the string theories would act like gauge field theories with gauge group $E(8) \times E(8)$ or $SO(32)$ respectively. Physicists jumped on the $E(8) \times E(8)$ theory because the exceptional group $E(8)$ contains $E(6)$ as a subgroup; and, as we know, $E(6)$ is an excellent candidate for a grand unified group. The $E(8) \times E(8)$ heterotic string, therefore, held out the promise of unifying gravity with the

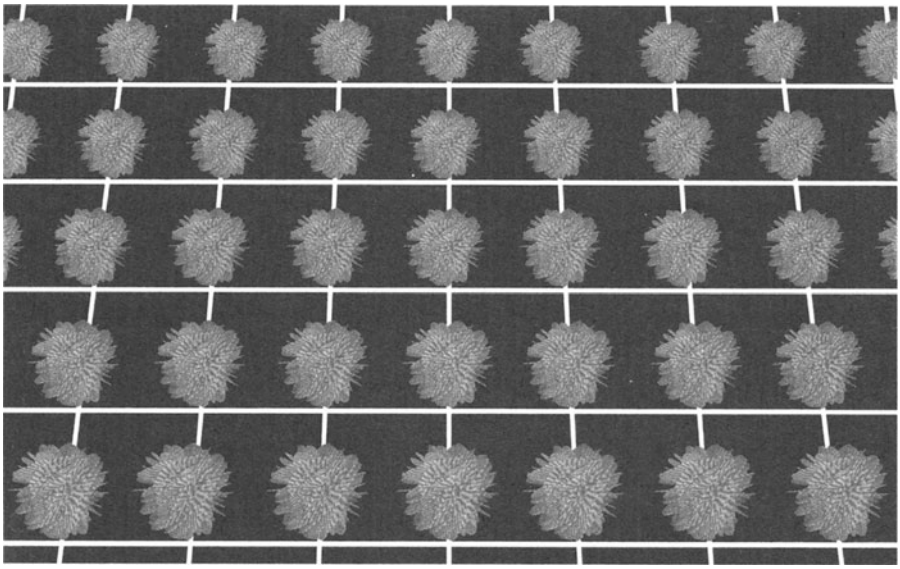


FIGURE 84 In superstring theory we need to imagine that there is a six-dimensional Calabi–Yau manifold at each point in space, just as in the original Kaluza–Klein theory we had to image that there was a circle at each point in space.

other three forces. Indeed, a few months after the paper by the Princeton String Quartet, Witten and three colleagues — Philip Candelas, then at the University of Texas, Gary Horowitz at the University of California, Santa Barbara, and Andrew Strominger, then at the Princeton Institute for Advanced Study — published a specific proposal for how the E(8) × E(8) heterotic string might make contact with the phenomena we observe at low energies. In their model, the particular Calabi–Yau manifold caused the E(8) × E(8) group to break down to the E(6) × E(8) group; at the same time, its $\mathcal{N} = 1$ supersymmetry was preserved. In other words, they had a four-dimensional effective theory with $\mathcal{N} = 1$ supersymmetry and a symmetry group of E(6) × E(8).

This discovery was perhaps the high point of the First Superstring Revolution. Starting with a string theory in ten dimensions, we end up with a supersymmetric grand unified theory (plus gravity) in four dimensions. There were other benefits with this model. For example, the compactification yields E(6) × E(8); the E(8) sector would describe matter that could interact gravitationally with the "E(6) matter" of the real world, but would be neutral under the E(6) gauge fields. In other words it would be unaffected by electromagnetism, the strong force, and the weak force, so we

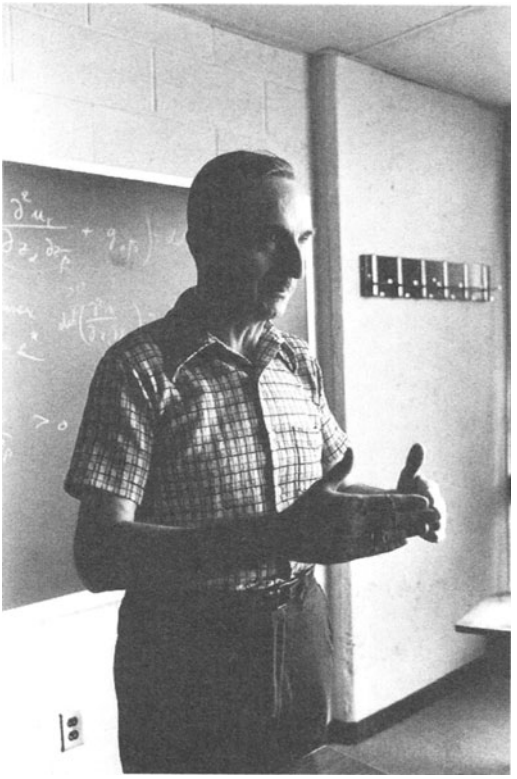


FIGURE 85 In 1954, the Italian–American mathematician Eugenio Calabi conjectured that a certain type of manifold would exist. It took mathematicians 22 years to prove that Calabi's conjecture was correct.

would never see it. A natural suggestion, then, was that this "E(8) shadow matter" comprises the dark matter that astronomers know is there in the cosmos but that physicists cannot readily account for. It gets better. By picking a particular topology for the Calabi–Yau manifold, one gets a prediction for the number of particle families: if a compactified Calabi–Yau manifold has three "holes" — in the same way that a doughnut has one hole — then we would observe three families of particles in our large dimensions! No other theory can explain so easily why we observe three families of particles. There are thousands of different Calabi–Yau manifolds (mathematicians have yet to tell us exactly how many), most of which predict a large number of particle families; but there are a few spaces that will produce the observed three families of particles. And the topology of those same spaces provide reasons for expecting the E(6) GUT group to break down to the SU(3) × SU(2) × U(1) of the Standard Model.

The E(8) × E(8) heterotic string, compactified on a six-dimensional Calabi–Yau manifold, looks like it might be *the* unified theory we have been looking for.

Well, as always, there are problems.

FIGURE 86 The Chinese–American mathematician Shing-Tung Yau is the recipient of the Fields Medal, as well as many other prestigious awards. In 1976, Yau proved a conjecture made in 1954 by Eugenio Calabi regarding the existence of a particular type of space. These special types of space are now known as Calabi–Yau manifolds.

The picture sketched above caused much excitement amongst physicists in 1985 and for good reason: it offered a true unified theory that came close to describing realistic physics. Indeed, some elements of the picture may even turn out to be correct. Almost twenty years later, however, we can point out the arbitrariness of the picture and we can identify the elements that do not work. Even back then, in the immediate aftermath of the First Superstring Revolution, several physicists pointed out what they believed were shortcomings of the theory.

Glashow and others, for example, were concerned that superstring theory was more akin to pure mathematics than to physics. The theory seemed to make no testable predictions and thus it could not be falsified; in this sense, the theory was not even wrong. (Mathematicians, who do not need to worry about whether their work accords with experimental observation, were keen to investigate the mathematical structures inherent in superstring theory, irrespective of its use in describing the real world. Such activity continues to the present day, with advances in mathematics being made as a direct result of superstring research.)

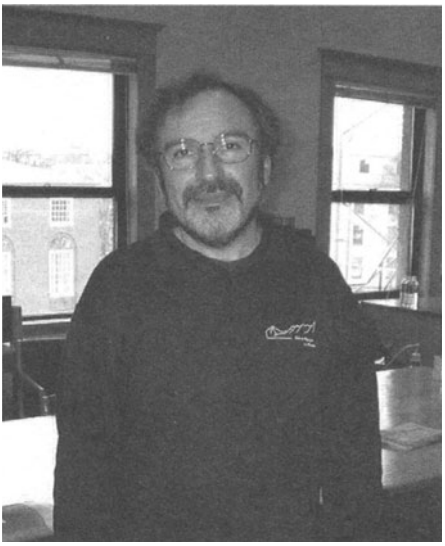


FIGURE 87 Philip Candelas is a professor at the Mathematical Institute, Oxford University. In 1986, while based at the University of Texas at Austin, he was a coauthor with Horowitz, Strominger, and Witten of an imaginative scheme that showed how heterotic strings might relate to observable phenomena.

FIGURE 88 Gary Horowitz is a professor of physics at the University of California, Santa Barbara. In 1986, with Candelas, Strominger, and Witten, he showed how heterotic strings might relate to observable phenomena. More recently, he has investigated how string theory can help us understand the physics of black holes.

In fact, specific superstring models do make predictions that might someday be tested. A common prediction, for example, is that there is another neutral vector boson — a particle like the Z^0, but more massive. Even if such a particle were found, however, it would not prove the case for strings: there are non-string scenarios in which such a particle could be accommodated. As physicists have learned more about string theory, it seems that the situation is not quite so gloomy as Glashow and others first thought. But 20 years ago, it seemed that the superstring idea would be forever untestable and even now the potential experimental tests are speculative.

Had we discovered superstring theory *first*, before all the rest of physics, then we would be mightily impressed with it. We saw earlier how writing down a consistent theory based on string ideas leads to an extremely rich

set of features. In fact, had we discovered superstring theory first, it would seem to make three basic predictions: gauge theories, general relativity, and supersymmetry. Unfortunately, we already know about gauge theories and general relativity, so we cannot take their existence as proof of the validity of superstring theory. These are, at best, postdictions of superstring theory. The experimental discovery of supersymmetry at the Tevatron or LHC would be an important boost for supporters of superstrings, but even then the case for superstrings would not be proven beyond reasonable doubt: nature might employ supersymmetry without necessarily making use of strings. So these features of superstring theory, whilst surely being encouraging enough for us to investigate the theory further, are not compelling enough for us to hail it definitively as the unified theory we seek.

Another difficulty for string theory was the arbitrariness involved in developing realistic models of the world we observe. A fundamental unified theory should let us calculate answers to fundamental questions, rather than have us measure certain values and feed them in to the theory by hand. For example, why should only three space dimensions be large? Why do we not observe four large space dimensions? Or five? Or, indeed, any number of dimensions up to the maximum allowed? Part of the difficulty in answering general questions such as these is that they cannot necessarily be addressed using the familiar perturbative approach. Superstring theory has a fundamental constant — the string coupling constant. If the constant is much less than one, we can ignore diagrams with lots of handles (the complicated diagrams appearing in Figure 78 on page 160) and quickly arrive at accurate answers. If the constant is greater than one, we can have no confidence at all in answers coming from a perturbative approach. Unfortunately, no one knows what value the string coupling constant takes in superstring theory. The constant is determined by the size of the dilaton field (we met the dilaton field when we discussed Kaluza–Klein theories). There is no particular reason why the dilaton field should be small — so it is unlikely that a realistic string model giving rise to our low-energy four-dimensional world could be analyzed accurately within a perturbative framework. In order to understand superstring theory better, in order to be truly convinced of its validity, we probably need an understanding of its nonperturbative aspects. No small task.

A further difficulty with superstring theory (which continues to this day) was the complexity of Calabi–Yau manifolds, the six-dimensional spaces upon which the theories are compactified. There are many thousands of Calabi–Yau manifolds, and each of them has a different topology. (They have holes, handles, and tunnels.) Suppose one could fix the topology, though; would that not simplify matters? Well, only a little. For a Calabi–Yau manifold with a particular topology, there are still hundreds of

parameters — called *moduli* — that describe the size and shape of the various compact dimensions. The Kaluza–Klein dilaton is an example of a modulus; it determines the compactification radius of the single extra dimension. Superstring theory has many such moduli, which would appear in our three-dimensional space as scalar fields determining local physical conditions. A set of moduli on a Calabi–Yau manifold generates a theory with particular values of parameters like the masses of particles and the coupling constants of forces — the sort of parameters that we want to explain in the Standard Model. We might hope that a particular set of moduli is singled out by the equations of string theory. Unfortunately, this does not happen: all values of the moduli give rise to a theory that is mathematically consistent. At present, we have no way of determining which set of moduli describes our Universe, nor why that particular set of values should have been chosen.

There are still other problems. For instance, after the most likely compactification of the extra six dimensions, we end up with a supersymmetric theory in four dimensions. That is fine, perhaps, but as we discussed in Chapter 5, we evidently do not live in a supersymmetric world. How and why is SUSY broken at low energies? String theory remains silent on this question. A related question has to do with particle masses: the lightest particles arising from string oscillations are massless in ten dimensions; after the compactification process, they give rise to lots of massless particles in four dimensions. But we know for certain of only two exactly massless particles: the photon and the gluon. The masses of the other particles, like the electron for example, might arise when SUSY breaks; but, as mentioned earlier, no one understands the details of that process.

Then there was the problem of the cosmological constant. The equations of general relativity contain a parameter — the cosmological constant — that describes the total energy of the vacuum. From astronomical observations it is known that the cosmological constant is small. String theory, being a fundamental theory of gravity, should be able to predict the value of this number. Well, according to string theory, if SUSY is unbroken then the cosmological constant is exactly zero. Contributions from bosons and fermions exactly cancel. Unfortunately, when SUSY breaks — as it must, since we do not observe supersymmetry in our everyday world — then string theory tells us that the cosmological constant develops a huge value: about 10^{100} times larger than the bound set by experiment. This is the worst discrepancy between observation and theoretical prediction in all of science.

For some people, the number of spacetime dimensions required by superstrings — precisely ten — was unsettling. As studies of supergravity had shown, nature seems to allow dimensions up to a maximum of 11.

Why should nature make use of ten dimensions rather than 11? It seemed strange.

Perhaps the major objection to superstrings, though, was aesthetic. There are *five* different theories that satisfy the many constraints imposed upon superstrings. Since we are searching for *the* unified theory that describes our world, this is an embarrassment of riches. If only *one* theory survived the constraints, if only *one* theory were singled out, then we would have strong grounds for hoping that this was *the* unique theory. But this does not happen.

For the record, the five consistent superstring theories are as follows.

Type I theory is a theory of unoriented, open strings (though, because the endpoints of an open string can touch, and join together to form a closed string, the theory describes both open and closed strings). As with the other superstring theories, the low-energy approximation of a type I string is a supergravity theory in ten dimensions. There is a massless graviton and several gauge fields that behave in many ways like the familiar photon, gluons, $W^{\pm}$, and Z^0 boson fields of the real world. There are fermion "matter" particles too. They are "charged" under the gauge fields (just as the electron is charged under the electromagnetic field, and the quark is charged under the color field), and the charges appear at the end of the open string. So, at least in ten dimensions, we have a low-energy theory that contains both gravity and gauge fields. The gauge symmetry of the type I string is SO(32), which is *very* much larger than the $SU(3) \times SU(2) \times U(1)$ symmetry of the Standard Model, so the theory would seem to predict too many forces and force carriers. The theory is parity violating (which is good) and contains $\mathcal{N} = 1$ supersymmetry (also good). In total, type I superstring theory makes a decent attempt at being a realistic unified theory. To string theorists of the First Superstring Revolution, the problems that we mentioned above might perhaps have appeared surmountable given sufficient time and ingenuity to apply to the difficulties.

Type IIA theory is a theory of unoriented, closed strings with $\mathcal{N} = 2$ supersymmetry. Vibrations that travel clockwise around the string are identical to those travelling anticlockwise; in other words, you cannot distinguish between right-movers and left-movers (it is a non-chiral theory). So, although the theory contains fermions, it would seem to fail as a theory of the real world, because the fermions may be either left- or right-handed. There is another problem with type IIA theory as a potential unified theory of physics: it has no gauge symmetry. So although it contains gravity (*all* the string theories possess a massless graviton) it does not contain the gauge symmetries that might give rise to the Standard Model.

Type IIB theory is similar to type IIA theory, except that this time the strings are oriented: one can distinguish between the right-movers and left-movers. This means that the theory is chiral, and that there will be left-handed fermions with no corresponding right-handed partners. Like the type IIA theory, however, there seems to be no way to add gauge forces into the picture.

Heterotic SO(32) theory is one of the two mathematically consistent hybrid string theories. It is an oriented, chiral, closed-string theory with $\mathcal{N} = 1$ supersymmetry. The gauge symmetry is SO(32) which, as we have already mentioned, is larger than the SU(3) × SU(2) × U(1) of the Standard Model. Upon compactification, the chirality that is present in the ten-dimensional theory can easily be destroyed. Once more, the theory would seem to fail as a theory of our four-dimensional world.

Heterotic $E(8) \times E(8)$ theory is the other mathematically consistent hybrid string theory. Except for the gauge group, it is similar to the SO(32) heterotic string. The difference in gauge group is important, however; this theory holds out the promise of giving us a realistic understanding of the physics we describe by the Standard Model. As we discussed earlier, after the First Superstring Revolution the heterotic E(8) × E(8) theory was strongly favored as *the* unified theory.

The five superstring theories, and their properties, are listed in Table 2.

TABLE 2 There are five distinct consistent superstring theories.

Type	Open/Closed	Oriented	$\mathcal{N}$	Details
I	Open (plus closed)	No	1	Graviton, no tachyon, SO(32) gauge symmetry
IIA	Closed	No	2	Graviton, no tachyon, no gauge symmetry
IIB	Closed	Yes	2	Graviton, no tachyon, no gauge symmetry
Heterotic SO(32)	Closed	Yes	1	Graviton, no tachyon, SO(32) gauge symmetry
Heterotic E(8) × E(8)	Closed	Yes	1	Graviton, no tachyon, E(8) × E(8) gauge symmetry

The five theories have some properties in common. For example, they all exist in a ten-dimensional spacetime. All of them contain the graviton. And all of them have a form of supergravity as their low-energy effective theory. (Interestingly, there are only *four* different ten-dimensional supergravity theories here: the E(8) × E(8) heterotic string, the type IIA string, and the type IIB string give rise to three different supergravity theories; but the SO(32) heterotic string and the type I string give rise to the *same*

supergravity theory. So, from the start, there are indications that the SO(32) heterotic string and the type I string might be linked.) Nevertheless, the five theories appear to be distinct, and the trouble is that each of these five theories is in some sense as fundamental as the others. We may prefer the heterotic E(8) × E(8) theory because it appears to generate more realistic low-energy physics, but that preference would hardly explain why nature allows, and yet avoids, four other possibilities. We want nature to single out *one* fundamental theory; we are looking for *the* unified theory. And here nature seems to be giving us *five* equally valid theories.

Five is four too many.

8

THE STORY OF M

Without the making of theories I am convinced there would be no observation.
Charles Darwin to Charles Lyell, *The Life and Letters of Charles Darwin*

About ten years after the First Superstring Revolution, physics saw the upheavals of the Second Superstring Revolution. The shot that started the revolution was fired by Witten on March 14, 1995, when he spoke at the annual gathering of the string community. Witten, building on recent work by several physicists, instigated a change in how we view strings that was perhaps even more important than what Green and Schwarz had proposed a decade earlier.

In the second revolution, physicists showed how the five different superstring theories are deeply related. In fact, the five theories are different manifestations of a single underlying theory — a theory that shows every indication of being *the* unique theory we have been searching for. The underlying theory is called M-Theory, where the "M" stands for meta, mystery, magic, mother, membrane, matrix, master, or whatever a particular author prefers. Unfortunately, this underlying theory is not well defined. Indeed, at present, there really is no "theory" that people can write down and investigate. Nor is there a grand principle from which the theory springs logically and fully formed. This situation is quite different to

how general relativity developed: Einstein had a beautiful idea — the principle of equivalence — from which his theory flowed naturally. Present-day physicists are reduced to searching for scraps of knowledge about the underlying theory and piecing the scraps together in order to formulate the theory with more precision. Once physicists have more of the theory written down, perhaps they will learn something about its guiding principles. Perhaps they will develop the M-theory equivalent of the principle of equivalence, and then they will better understand *why* the Universe is the way it is. But, in all honesty, it must be said that this seems far off.

Despite the mist and general murkiness in which M-theory is presently mired, it seems certain to be a topic of interest for many years — perhaps decades — to come. When physicists finally do write down the theory, they may well have written down the final theory of physics. But that would be the endpoint of the story; the beginning of the story relates, of course, to symmetry and to an old concept called *duality*.

Electromagnetic Duality

The yin–yang principle . . . an explicit duality expressing an implicit unity.
Alan Watts, *Tao: The Watercourse Way*

The notion of duality is an old one in physics. It goes back at least to the 1860s and James Clerk Maxwell's discovery of the equations for the behavior in a vacuum of electric and magnetic fields. The Maxwell equations possess many beautiful symmetries, one of the most interesting of which is a symmetry under the interchange of the electric and magnetic fields. In particular, if you interchange the electric field with the magnetic field, whilst simultaneously interchanging the magnetic field with the negative of the electric field, the Maxwell equations are unaltered. This symmetry, an invariance of the vacuum electromagnetic equations under a transformation that swaps electric and magnetic fields, is called duality.

The Maxwell equations do not just govern the behavior of electric and magnetic fields in a vacuum. They apply to static charges. They also apply to moving charges, or currents. It turns out that duality holds in this general case, but only if we consider both electric *and* magnetic charges and electric *and* magnetic currents.

Electric charges we know about. Magnetic charges are more mysterious. They are commonly known as *magnetic monopoles*, but we have no experience with these objects. The bar magnets we play with as children always have a north and south pole: bar magnets are *dipoles*. If you cut a bar magnet in half, for example, you do not get two isolated north and south monopoles. Instead, you get *two* smaller dipole magnets. You can

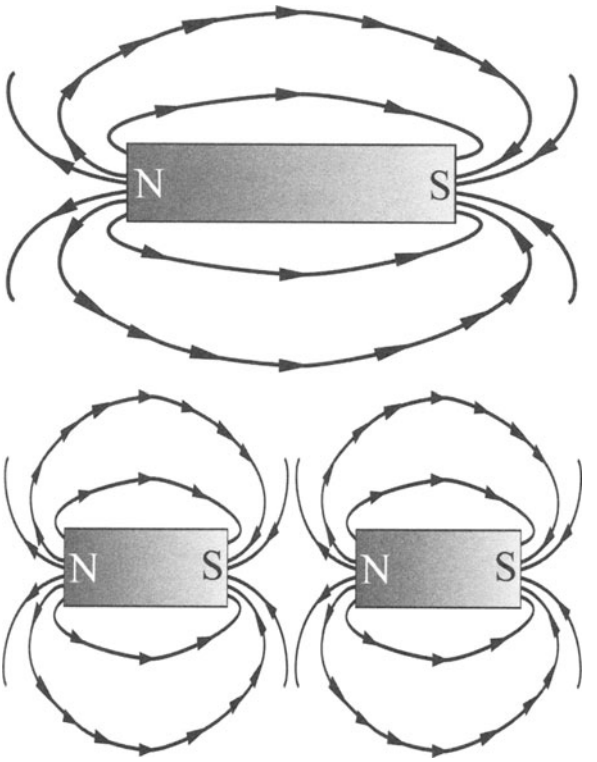


FIGURE 89 The bar magnet in the upper part of the diagram has a north pole and a south pole. The magnetic field lines loop continuously. Cut the magnet in half and we get two *magnets,* both *of which have a north and south pole, as illustrated in the lower part of the diagram. Isolated magnetic monopoles cannot be produced by cutting a bar magnet in half, and no other type of experiment has ever produced a monopole. It seems that they do not exist.*

keep on cutting a bar magnet in half, but you never obtain an isolated monopole. Instead, you eventually reach the protons, neutrons, and electrons that make up the magnet; these particles all possess a spin and are themselves tiny dipole magnets. You simply cannot cut an electron in half in an attempt to obtain a monopole.

Figure 89 shows the field lines of a dipole. You can actually see the magnetic lines of force if you sprinkle iron filings close to a bar magnet. The field lines loop continuously, leaving the north pole of the magnet, re-entering the magnet at the south pole, and then returning to the north pole through the body of the magnet. This is in marked contrast to the field lines around an electric charge; the electric field lines emanate from, or converge upon, the charge (see Figure 90). If the simple measure of cutting magnets in half does not produce magnetic monopoles, neither does the more sophisticated procedure of colliding particles together in high-energy accelerators; magnetic monopoles have never been observed in such experiments. Nor have monopoles been seen in cosmic ray experiments — or in any other type of experiment. The duality symmetry would thus seem not to hold in nature, because although we observe electric charges we do not observe magnetic charges.

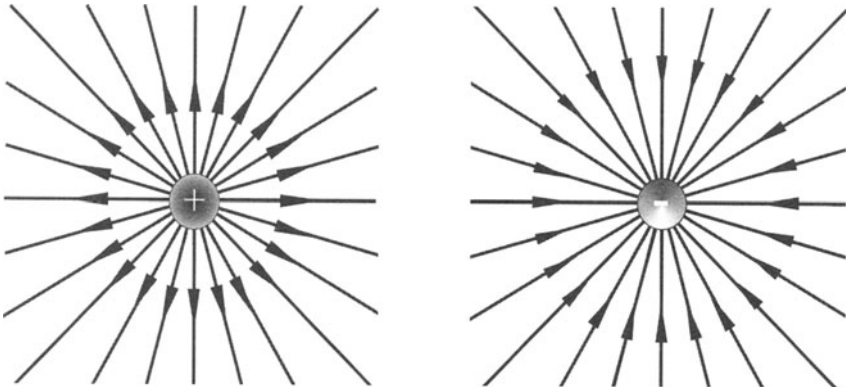


FIGURE 90 Unlike magnetic poles, electric poles — or charges — appear in isolation. The electric field lines either emanate from, or end on, an electric charge.

When Maxwell developed his ideas of electromagnetism it made sense simply to banish magnetic monopoles from the theory — if monopoles do not exist then it would seem that nothing is lost by ignoring them. And Maxwell found that his equations were simplified: he had to deal only with electric charges, the movement of which generated magnetism. Magnetic monopoles were thus absent from the equations of classical electromagnetism — and therefore absent from the quantum theory that grew out of electromagnetism. But by ignoring monopoles we *do* lose something: the beautiful duality symmetry.

The return of monopoles

Paul Dirac wondered if there were a way of introducing magnetic monopoles into quantum physics without losing all those successes that came from assuming monopoles did not exist. In 1931, in a wonderfully imaginative paper, he showed that he *could* successfully introduce monopoles into the theory — but only if the magnitude of the basic magnetic charge was linked to the magnitude of the basic electric charge. He showed that, in appropriate quantum units, the product of the basic electric and magnetic charges — in other words, the product of the electron charge and the monopole charge — must be an integer. (If the product of two numbers is kept constant, we cannot vary them in such a way that they both increase. If one of the numbers becomes large, the other must become small. So in Dirac's theory if the basic magnetic charge takes a large value the electric charge must be small, and vice versa. Of course, we know the value of the basic electric charge: it is the charge on the electron. This fixes the numbers in the theory. It turns out that the basic magnetic charge, in appropriate units, is 68.5 times larger than the electron charge.)

It had always been a mystery why we only ever see electric charges that are integer multiples of the charge on the electron. (Quarks possess fractional charge but they are never seen in isolation. They always combine in such a way that the objects we actually observe possess an integer multiple of the electron charge.) As long as magnetic monopoles exist, Dirac had solved the mystery. His theory showed that if just one magnetic monopole exists somewhere in the Universe, all isolated electric charges must be integer multiples of the reciprocal of the magnetic charge. In other words, electric charge must be quantized.

Despite the undoubted elegance of Dirac's result, little research was undertaken on magnetic monopoles for more than 40 years. The next advance did not come until 1974 when Gerardus 't Hooft, at Utrecht University, and Alexander Polyakov, then at the Landau Institute for Theoretical Physics in Moscow, independently showed that the existence of monopoles is inevitable in a wide class of grand unified theories.

Polyakov and 't Hooft showed that, when the gauge field symmetry of a grand unified theory breaks, the fundamental fields can become "twisted." A "defect" forms in the fields. The solutions of the field equations allow for bound, twisted heaps of field energy. In certain configurations, these heaps look like the quantum particles we have met so far — they have a mass, spin, charge, and so on — but they are *not* the quanta of fundamental fields in the way that quarks, leptons, and gauge bosons are. (They probably do not fit your mental picture of a quantum particle. They are spread out in space and are *extremely* massive, maybe 10^{18} times as massive as the proton.) It is these extended, twisted configurations of fundamental fields that 't Hooft and Polyakov identified as magnetic monopoles: far away from the heap of field energy, all we see is a radially directed magnetic field. (Similarly, when we are far away from an electric charge, all we see is a radially directed electric field.)

The type of field configuration that 't Hooft and Polyakov investigated is known as a *soliton*, a phenomenon that has been a curiosity for many decades. The first example of a soliton was noticed in August 1834, when John Scott Russell saw a "solitary wave" in the Union Canal near Edinburgh. It was quite unlike the behavior of any waves Scott Russell had seen at the beach. The wave was a "rounded, smooth, and well-defined heap of water," which travelled at about 8 miles per hour for a couple of miles without changing its speed or its shape. We now know of other instances of solitary waves. One of the most famous examples of a solitary water wave is the Severn bore in England (see Figure 92), which travels for about 40 kilometers and can reach heights of 2 meters. People have been known to surf the bore for up to 10 kilometers, riding one continuous wave in the middle of the English countryside.

FIGURE 91 In August 1834, John Scott Russell watched a boat being drawn along the Union Canal by a pair of horses. When the boat suddenly stopped, he noticed that the bow wave continued rapidly forward for a couple of miles. It was a soliton. In 1995, volunteers recreated the famous solitary wave observed by Scott Russell.

A soliton is now defined as a localized disturbance in a continuous medium that can propagate over long distances with no change to its shape or amplitude, even if two or more such disturbances collide. The heaps of field energy in the theories of 't Hooft and Polyakov follow this definition.

Soliton Technology

Solitons are important in many branches of science and technology, not just field theory. Billions of solitons per second now carry information down the fibre-optic circuits of our networked world. And solitons find applications in fields as diverse as high-temperature superconductivity and the modelling of energy transport in the DNA molecule.

A reasonable question to ask is: what causes the 't Hooft–Polyakov soliton to persist? Why does it not dissipate into its constituent fields, just as the solitary wave of Scott Russell eventually diminished in the windings of the Union Canal? Consider the configuration of fields in Figure 93. The field configuration on the left, in which all the arrows point in the same direction, corresponds to the vacuum — the state of minimum energy.

FIGURE 92 Where a river estuary is the correct shape and the tidal conditions are just right, a bore — a "solitary wave" — can form. The requirements for the formation of a bore are not often met and there are only about 60 tidal bores in the world. The Severn bore in England, shown here, is one of the best-known examples; records of this bore date back as far as the sixth century.

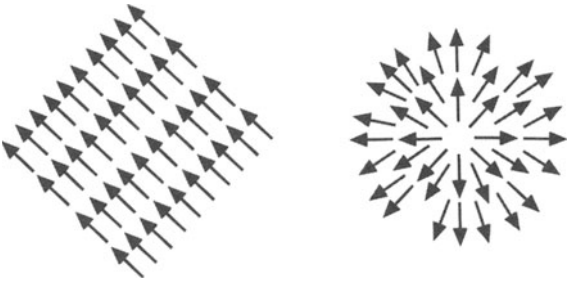


FIGURE 93 On the left, the field lines all point in the same direction. This configuration of field lines corresponds to the vacuum, or the state with lowest energy. On the right, all the arrows point radially outward. This configuration possesses more energy, and so would be expected to decay into the vacuum configuration. However, it is impossible to move smoothly from the configuration on the right to the configuration on the left; so although the set of field lines on the right possesses more energy than the vacuum state, it is a stable configuration. Topological considerations thus prevent the hedgehog decaying into the vacuum.

Polyakov and 't Hooft investigated the field configuration on the right, in which all the arrows are directed radially outward; it is called the hedgehog configuration (it looks like the spines on a curled-up hedgehog). When we ask why the soliton does not dissipate we are asking why does the hedgehog configuration not eventually decay into the vacuum configuration? This is the sort of question that can be investigated by topologists.

Topology is the study of objects undergoing continuous transformations. (A continuous transformation means that you can pull, prod, stretch or deform the object as much as you wish; but you are not allowed to puncture or tear the object in any way.) So the surface of a perfect sphere and the surface of a potato are topologically identical: you can continuously transform a sphere into any potato, a cigar, a box, or many other shapes. And a doughnut and a coffee cup are topologically identical: you can continuously transform the doughnut, with its hole, into a coffee cup. But the sphere and the doughnut are topologically distinct: the doughnut has a hole and the sphere does not. Homer Simpson once drooled: "Doughnuts — is there nothing they can't do?" Well, yes: doughnuts cannot continuously transform into a sphere. In order to transform a doughnut into a sphere you have to first tear it, then stretch it, and then stitch it back together. Such an operation is not allowed in topology. The same consideration applies to the hedgehog and vacuum configurations. They are topologically distinct. The hedgehog is an example of a topological *defect*. In order for the soliton to dissipate, it would have to be cut and then stitched back together. Topology forbids that from happening. The 't Hooft–Polyakov monopole is thus a *topological soliton*. In a sense, we can think of the soliton as possessing a topological "charge," which is conserved just as electric charge is conserved. The only way for a topological soliton to be destroyed is if it meets a topological antisoliton.

The 't Hooft–Polyakov monopoles form automatically and in large numbers when the symmetry of a grand unified theory is spontaneously broken. So if the Standard Model is indeed unified through a GUT at high energies, as we suspect it should be, then the early Universe should have produced monopoles in abundance. Topology tells us that these monopoles should still be around. That we do not observe these relics from the Big Bang might therefore seem to be something of an embarrassment. However, the grand unified theories that imply the existence of monopoles also suggest that the very early Universe may have undergone an incredibly rapid period of expansion known as inflation (see the boxed text on page 123). Such inflationary expansion would have drastically diluted the density of monopoles — so much so that the observable Universe might contain just one monopole. If inflationary models are correct, then it is small wonder that we have not seen a monopole. So despite the lack of observational evidence for them, it is certainly worth examining theories in which monopoles exist, and one thing such theories can possess is a duality symmetry between electric and magnetic fields.

In 1974 't Hooft and, independently, Stanley Mandelstam of the University of California, Berkeley, proposed that a type of electromagnetic duality might explain one of the key puzzles of QCD: quark confinement.

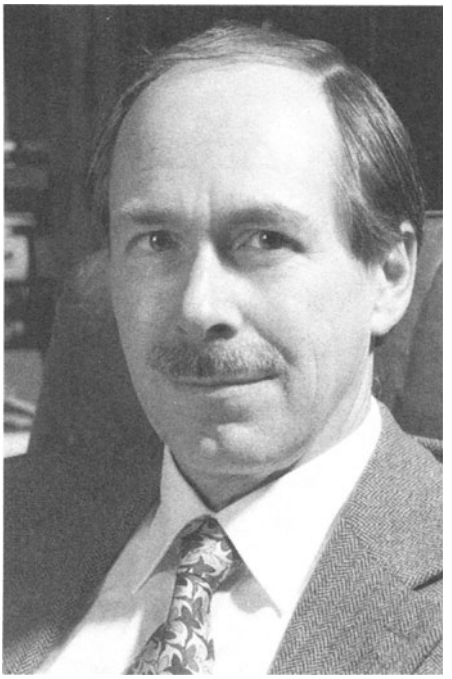


FIGURE 94 The Dutch physicist Gerardus 't Hooft won the 1999 Nobel Prize for physics, along with his colleague Martin Veltman, for his role in elucidating the quantum structure of electroweak interactions. He has also made important contributions in several other areas of theoretical physics, including his suggestion that grand unified theories in general imply the existence of magnetic monopoles.

As we know, quarks in standard QCD form bound states of mesons (a quark–antiquark pair) and baryons (a triplet of quarks). Physicists had no proof that confinement must happen, but both experiment and numerical simulation suggested that the energy contained in a quark–antiquark pair, say, grew linearly with the separation between the quarks. Double the separation between quark and antiquark, and the energy doubled. Quadruple the separation and the energy grew by a factor of four, and so on. In this scenario, we could never impart enough energy to separate quarks by a large enough distance to have an isolated, free quark.

A simple picture explains why the energy between quarks grows in this way: a "flux tube" forms between the quarks. The flux tube is a thin cigar-like volume containing the field lines of the color force, and it acts as if it were a piece of elastic between the quarks. (See the left-hand diagram in Figure 95.) It is because the energy grows linearly in this flux tube that one cannot separate a quark–antiquark pair by an infinite distance and study a quark in isolation. Quarks are confined.

The difficulty for theorists was that this picture takes place in the strong coupling regime, a place where the usual theoretical tools of the perturbation approach do not work. Physicists had to resort to computer simulations in order to try and study confinement; analytic results from theory were simply impossible to come by. However, there is another phenomenon in

FIGURE 95 When we try to separate a quark–antiquark pair, the color lines of force between them become stretched, and a flux tube forms between the quarks. Something similar happens if we insert a monopole–antimonopole pair into a superconductor; the magnetic field lines are confined to a flux tube.

physics that is similar to quark confinement but that is much more down to earth and much more accessible: the creation of a magnetic flux tube between monopoles in a superconductor (see the boxed text on page 191).

There is a clear analogy here. Quarks carry color charge and they are confined in a vacuum. Monopoles carry magnetic charge and they are confined in a superconductor. Remembering that color charge is the QCD equivalent of electric charge, perhaps the QCD vacuum bears the same relationship to a superconductor as electricity bears to magnetism! This idea of 't Hooft's was an interesting application of duality, and it explained in a qualitative way why quarks were confined. Unfortunately, in the 1970s neither 't Hooft nor anyone else had the necessary tools to develop the idea.

The 1970s saw one more development of duality and the monopole idea. In 1977, Claus Montonen and David Olive, who were both then working at CERN, considered a particular supersymmetric quantum theory containing monopoles and conjectured that the theory had an exact symmetry between electric charge and magnetic charge. You could swap "electricity" with "magnetism" and nothing would be altered. The two descriptions would be dual to each other. This electromagnetic duality, if it existed, might seem surprising after what we have learned about monopoles. We normally think of objects carrying electric charge (quarks, electrons, and so on) as being light, point-like, elementary particles; objects carrying magnetic charge (magnetic monopoles) are heavy, spread-out, extended solitons. But in the dual formulation, it is the monopoles that are elementary particles, and the quarks and electrons that are solitonic heaps of energy. In other words, an elementary particle with a particular electric charge would be equivalent to a solitonic particle possessing the reciprocal of that charge. Swapping electric and magnetic charges causes other things to be swapped. For example, the coupling strength of the theory depends upon the value of the electric charge. So in the Montonen–Olive theory, swapping a coupling constant that depends on the charge with a coupling constant that depends on the reciprocal of the charge should have no effect on physics. Now, if the coupling constant is small (so we are in the weak coupling regime) then the reciprocal of the coupling constant is large (so we are in the strong coupling regime). There is thus a conjectured duality between the weak and strong coupling regimes of the theory.

Monopoles in a Superconductor

In Berlin, in 1933, Walther Meissner and his graduate student Robert Ochsenfeld observed that a superconductor expels magnetic flux. This is now called the Meissner effect.

Monopoles are a source of magnetic flux, so what happens when we insert a monopole–antimonopole pair, separated by a large distance, into a superconductor? What happens when the superconductor tries to expel the flux? (In practice, of course, we cannot work with monopoles. But it is possible to replicate this gedanken experiment by applying an appropriate combination of external magnetic fields.) The answer is that the system forms a configuration that possesses the least possible energy: a nonsuperconducting tube forms between the monopole–antimonopole pair, and the magnetic flux is confined to this tube. See the right-hand diagram in Figure 95. (The tube is called an Abrikosov–Gorkov flux tube, after the Russian physicists Aleksei Abrikosov and Lev Gorkov, who first described the phenomenon. Abrikosov won the 2003 Nobel Prize for physics for his pioneering work in superconductor physics.) The Abrikosov–Gorkov flux tube has a certain energy per unit length. So if you double the separation between monopole and antimonopole, the energy doubles; quadruple the separation, and the energy grows by a factor of four; and so on. This should sound familiar!

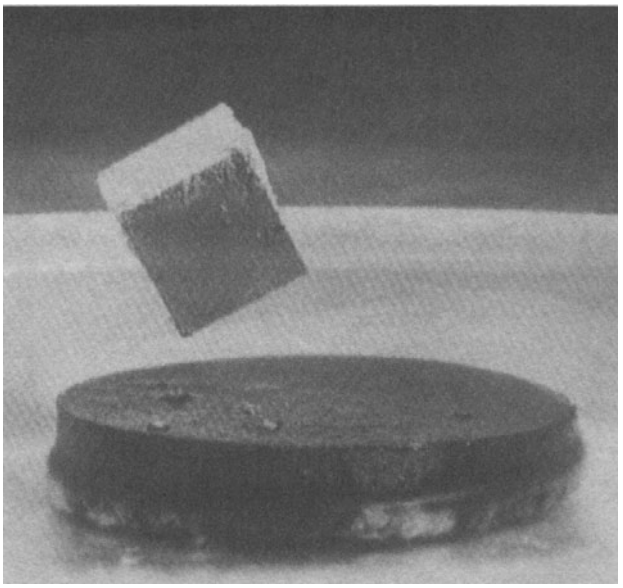


FIGURE 96 The classic demonstration of the Meissner effect is when a magnet (above) is brought close to a superconductor (the disk below): the magnet encounters a repulsive force and can be made to levitate. This property has practical applications in technologies ranging from high-speed magnetically levitated trains to MRI scanners (magnetic resonance imaging, or MRI, is now routinely used to examine almost all organs of the body).

Montonen and Olive were interested in grand unified theories. If such theories really did possess electromagnetic duality, a new way of studying their non-perturbative aspects opens up. Suppose you come across an unsolvable problem in the strong coupling regime. The perturbative approach does not work and progress appears to be impossible. Well, you merely have to move across to the dual theory, which is weakly coupled, and you can solve the dual problem using your usual perturbative techniques. Since the theories are dual, they describe the same physics and the final result is the same in either theory. What is impossible to solve in one theory is easy to solve in the dual theory, and vice versa. What a wonderful trick!

Unfortunately, Montonen and Olive could not *prove* that electromagnetic duality existed in their theory. The reason is obvious: their proposed duality exchanges weak coupling with strong coupling, but we only really understand quantum field theory in the weak coupling regime. Since it is impossible for both the coupling constant and its reciprocal to be simultaneously small, one of the dual pair of theories is out of reach. The strong coupling regime remains hidden from us, and we simply cannot know for certain whether duality holds. In the absence of ways of testing the Montonen–Olive conjecture, the whole subject of duality went into decline.

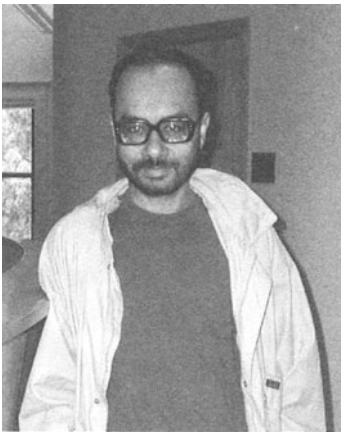


FIGURE 97 Ashoke Sen is a professor of physics at the Harish-Chandra Research Institute in Allahabad, India. He has made several important contributions to string theory.

It was not until the 1990s that physicists began to find good evidence for the existence of electromagnetic duality in their theories. Ashoke Sen showed that supersymmetry demanded the existence of a certain type of soliton with both magnetic and electric charges, something that Montonen and Olive had conjectured but could not prove. The discovery inspired physicists such as Nathan Seiberg of Princeton University to take more seriously the notion of duality in supersymmetric field theories. It turns out that, if

a theory is supersymmetric, one can deduce several pieces of information about the strongly coupled regime. Supersymmetry — as it now appears — is the key to unlocking the secrets of strong coupling. For example, there exists a type of quantum state called a BPS state (named after the three physicists who, in the mid-1970s, first investigated its properties: Eugene Bogomolny, then at the Landau Institute in Moscow, and independently Manoj Prasad and Charles Sommerfield, then at Yale University). Symmetry considerations ensure that a BPS state possesses, for a given charge, the minimum possible energy — and supersymmetry ensures that a BPS state maintains its minimum possible energy, regardless of the coupling strength of the theory. So we can calculate the energy of a BPS state in the weak coupling regime and, since supersymmetry "protects" the state, extrapolate this information with confidence to the strong coupling regime. We have a glimpse of the non-perturbative region of the theory. The spectrum of BPS states is by no means the complete spectrum of particles in the strong coupling regime. But combined with other pieces of information coming from quantum mechanics, one can often identify a distinctive pattern — a "thumbprint" — in the strongly coupled theory. From this thumbprint, one can then deduce the dual theory.

Seiberg followed this kind of logic to produce several examples of duality, and his work provided overwhelming evidence in favor of the idea that supersymmetric gauge theories possess electromagnetic duality. In strict terms, perhaps, this remains circumstantial evidence, but it would be strong enough evidence to obtain a guilty verdict in most courts. Few people now seriously doubt that the sorts of theories that we have seen might describe physics at the GUT scale — supersymmetric gauge theories — possess duality.

In 1994, Seiberg and Witten published a paper in which they studied a four-dimensional theory with $\mathcal{N} = 2$ supersymmtry and an SU(2) gauge symmetry. Until then, the best that physicists could have hoped to gain from such a theory was a set of perturbative approximations. But using the tremendous power of duality, Seiberg and Witten were able to find *exact* results in the strong coupling regime of the theory. The results were soon extended to general SU(N) gauge groups. This was the first time that exact results had ever been found in such field theories. Indeed, only a few months before the publication of this paper, many physicists had doubted whether exact results could *ever* be obtained for a four-dimensional theory! One of the key findings of Seiberg and Witten was a *quantitative* description of quark confinement. This was neither a hand-waving argument, nor numerical evidence from computer simulations. This was an analytic result. We now know, thanks to Seiberg and Witten, that $\mathcal{N} = 2$ supersymmetric

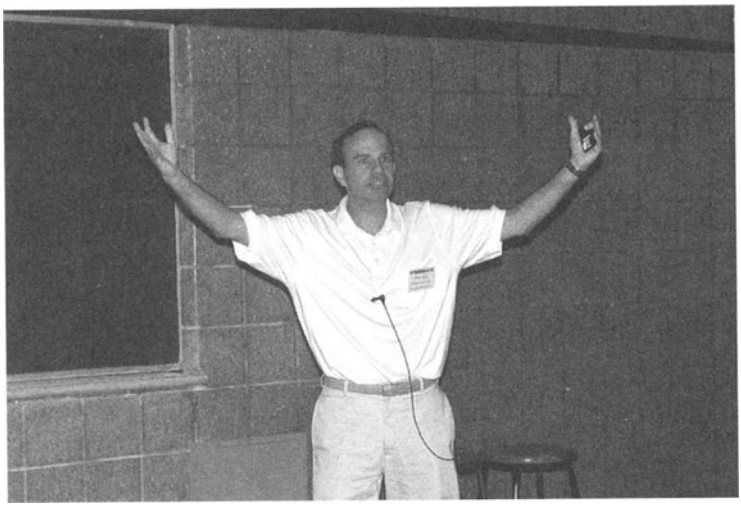


FIGURE 98 Nathan Seiberg is a professor of mathematical physics at the Institute of Advanced Study in Princeton. He has collaborated with Ed Witten on several groundbreaking papers.

QCD definitely exhibits confinement. (The mechanism is by monopole condensation, as suggested in the 1970s by 't Hooft.)

Of course, supersymmetric QCD is not the QCD of the Standard Model. As we have stressed, supersymmetry has not yet been observed in nature. So the proof of confinement given by Seiberg and Witten was *not* a proof that quarks are confined in ordinary QCD. Nevertheless, at the very least, the Seiberg–Witten approach made quark confinement in ordinary QCD seem more plausible.

Duality and Mathematics

The Seiberg–Witten approach turned out to be a gift for those pure mathematicians who study the topology of smooth four-dimensional manifolds. Using duality, mathematicians were able to prove many results that were previously intractable. Although physicists claim Witten as one of their own, mathematicians are more than happy to use the fruits of his work.

With both physicists and mathematicians suddenly enamored with the delights of duality, it is perhaps not surprising that they began to look again at superstrings. The key to proving duality is supersymmetry; since string theory makes sense only if supersymmetry holds, could duality be found here, too?

Duality in Strings

When such strings jar, what hope of harmony?
William Shakespeare, *Henry VI, Part 2*

Duality has indeed been the key to unlocking many of the secrets of string theory. Witten's talk at the *Strings '95* conference crystallized the work on duality made by several string physicists — and gave the world a glimpse of a surprising, but beautiful, underlying theory.

S-duality

Two of the physicists who inspired Witten were Christopher Hull of Queen Mary, University of London, and Paul Townsend of Cambridge University.

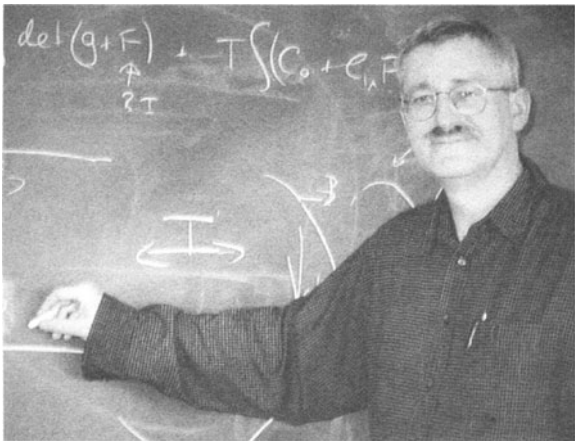


FIGURE 99 Paul Townsend is a professor of theoretical physics at Cambridge University. He was one of the physicists who coined the word 'brane,' and was one of the first to recognize the importance of duality in string theory.

Hull and Townsend looked at the behavior of type IIB string theory and investigated what happens when its dilaton field is transformed in various ways. They found the theory was unchanged if they replaced a positive value for the dilaton field (ϕ, say) with a value that had the same magnitude but was negative ($-\phi$). In other words, swap ϕ with $-\phi$ and you have two identical theories. There is a duality between them. Remember, though, that the coupling constant of a string theory depends upon the size of the dilaton field in a particular and well-defined way. When we replace a positive value for the dilaton with a corresponding negative value, we are exchanging a large coupling constant with a small coupling constant. This symmetry between the strong coupling regime and the weak coupling regime of a theory is called *S-duality*. (We can think of S-duality as being short for "strong–weak coupling duality.") Type IIB string theory is thus self-dual. This is wonderful news, because it means we can understand the

strong-coupling regime of the type IIB theory, which is difficult to work with, by studying the weak-coupling regime of the theory, which is easy to work with — inasmuch as anything is easy in string theory.

Witten then looked at type I string theory and SO(32) heterotic string theory. Since both of these theories have SO(32) as a gauge group, perhaps they were related by duality? He found that if he exchanged a positive value for the dilaton of type I string theory with a negative value for the dilaton of SO(32) heterotic string theory, nothing changed. So Witten conjectured that the type I and SO(32) heterotic string theories are related by S-duality. When the coupling constant of one theory becomes large, it can be thought of as transforming into the other theory. A duality like this seems almost magical, since at first glance the two theories are totally unrelated. Yet in a sense they are the same theory.

S-duality not only lets us understand the strong-coupling regime of type I, type IIB, and SO(32) heterotic theories, it reduces the number of distinct superstring theories from five to four: we need not think of type I and SO(32) heterotic theories as being fundamentally different.

It gets better.

T-duality

As long ago as 1984, Keiji Kikkawa and Masami Yamasaki discussed the possibility of a string theory duality in the weak-coupling regime. Since Kikkawa and Yamasaki were studying the weak-coupling regime they could apply perturbation theory — which was what enabled them to discover this duality before S-duality. It turns out, however, that their duality also often holds in the strong-coupling regime. The duality now goes by the name of *T-duality*, which is short for "target-space duality."

The target space of a string theory is simply the spacetime in which we place it. The usual place for superstring theories is a flat ten-dimensional spacetime that is infinite in all directions. But we can also place the theories in a spacetime in which some of the dimensions have been compactified. Let's look at the simplest such case.

Consider a ten-dimensional spacetime (the usual combination of nine space dimensions plus one time dimension) and suppose that one of the space dimensions is compactified on a small circle. This is rather like the Kaluza–Klein geometry we looked at earlier. If the circle has radius R, a point particle can travel around the circle for a distance $2\pi R$ before it returns to its starting point. According to the rules of quantum mechanics, the momentum and energy of the particle is quantized. In fact, the energy of a particle is proportional to n/R, where n is an integer; so we get an infinite tower of particles with energies proportional to $1/R$, $2/R$, $3/R$,

and so on. Exactly the same thing happens when a string travels around the compactified circle: its momentum is quantized. The motion of a string around a compactified circle thus gives rise to *momentum modes*, just as in the point-particle case. In the point-particle case, that is the end of the story, but things are more complicated when we consider closed strings. In addition to moving in the direction of the circle, we can wrap a closed string *around* the circle. The number of times that a string is wound around a circle is called the *winding number.* Since the string is closed, we can wrap it around a circle only a whole number of times (the winding number is quantized, in other words). If the string is wrapped once around the circle, the winding number is 1. If it is wrapped twice around the circle, the winding number is 2, and so on. This configuration of strings, which cannot occur with point-like particles, gives rise to *winding modes.* Now, the more we stretch a string to wind it around the circle, the greater the string's energy; in other words, winding modes contribute to the energy of the string. In fact, if the string is wrapped m times around the circular dimension, the energy is proportional to $m \times R$; so we get an infinite tower of excited strings, with energies proportional to R, $2R$, $3R$, and so on. (See Figure 100.)

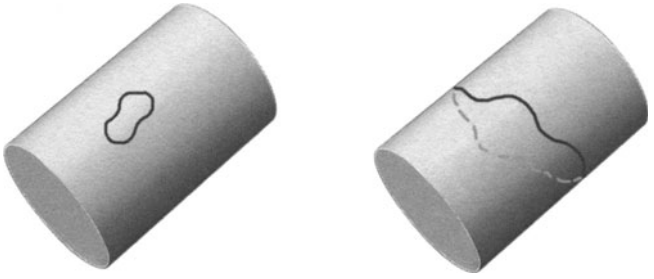


FIGURE 100 A string can move around a compactified circle, as in the left-hand side of this diagram. The behavior of the string in this case is similar to that of a point particle, and it gives rise to momentum modes of the string. A string can also wrap around a compactified circle, which is something that a point particle cannot do. This behavior gives rise to winding modes of a string: the more times the string is wrapped around the circle, the more energy the string possesses.

Now interchange the momentum modes, which have energies given by n/R, with the winding modes, which have energies given by $m \times R$. At the same time, exchange the radius R of the circle with the quantity L^2_{string}/R, where L_{string} is the fundamental string length. (If R is very much smaller than L_{string}, then L^2_{string}/R will be a very large distance. So we are swapping a small distance with a large distance.) If physics is identical in the two cases, we have two theories that are T-dual.

It turns out that type IIA string theory is T-dual to type IIB string theory. In other words, if we take these two theories and compactify them both on a circle, then swap the momentum and winding modes of the string while swapping a small distance scale with a large distance scale, we find that type IIA theory changes into type IIB theory (and vice versa). T-duality similarly relates the heterotic SO(32) theory with the heterotic E(8) × E(8) theory. The two type II superstring theories and the two heterotic theories are thus intimately related. We might start out with five distinct string theories, but T-duality reduces this number to just three. And since the type I string is S-dual to the SO(32) heterotic string, the number of different superstring theories is just two. In fact, as you can probably guess, it turns out that duality symmetries can link *all* superstring theories. *There is just one theory.*

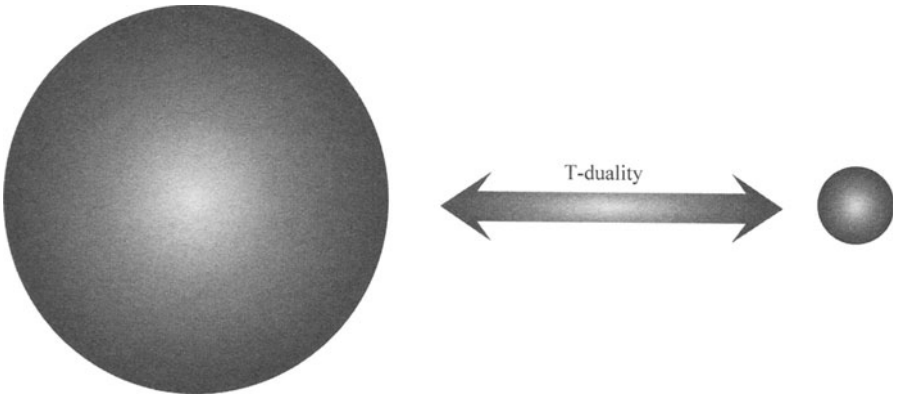


FIGURE 101 T-duality inverts the compactification radius. Physics with a large compactification radius is indistinguishable from physics with a small compactification radius.

T-duality is so counterintuitive — it says, in essence, that we cannot distinguish between a very large distance and a very small distance — that it is worth looking at it in a little more detail. After all, *of course* we know whether a compactified circle is large or small: ultimately, can we not just directly measure the circumference of the circle and thereby deduce its radius?

Our observable Universe extends for about 13.7 billion light years, and it is possible (though perhaps unlikely) that the three spatial dimensions curve back on themselves. In other words, it is possible that the three large spatial dimensions we observe are actually circles with a radius of about 13.7 billion light years. If our Universe were described by a T-dual string theory, then the physics we observe in our large Universe would be quite

T-duality and Topology

In string theory there can even be an ambiguity in the underlying topology of a spacetime. When we discussed T-duality we considered the simplest case of a single circular dimension; since only the size of the circle was altered, the topology was unchanged by T-duality. Brian Greene and Ronen Plesser have shown that, for realistic compactifications on Calabi–Yau manifolds, it is impossible to tell on which of a pair of *mirror manifolds* the theory was compactified. There can thus be a *mirror symmetry* between Calabi–Yau manifold pairs. (These mirror manifolds are not actual mirror images of each other; they can have quite different appearances and properties.) The physics of a string theory compactified on these two radically different manifolds is identical.

The property of mirror symmetry is important not only to physicists, but also to mathematicians. It turns out that problems that are intractable on one manifold are relatively easy to solve on the mirror manifold. Using techniques developed in string theory, mathematicians have been able to solve problems that previously they had not even thought to attempt. Once more, we see the close link between string theory and mathematics. At present it may be impossible to falsify string theory through experiment (which has been the usual method of progression for science), but mathematics is proving to be a useful testing ground of many ideas in string theory. So far, mathematicians have found no errors.

indistinguishable from the physics in a Universe with three circular dimensions of radius 10^{-96} meters. And as our Universe expands, the dual Universe would get smaller. But surely this is crazy! A Universe that has a radius of 10^{-96} meters — a minuscule fraction of the Planck length — is too small to accommodate a single electron, let alone the planets and the stars and the galaxies that we observe. And anyway, isn't the Planck length supposed to be the smallest distance we can meaningfully talk about? What sense can we make of dimensions that are so incredibly smaller than the Planck length? Does all of this not disprove the notion of T-duality?

Well, no it does not, as the boxed text on page 200 explains.

T-duality cannot occur in the physics of point particles, because points cannot wrap themselves around a circle in the way that strings can do. Getting wrapped around a circle is stringy behavior. If nature really is described by string theory, then certain fundamental conclusions would seem to follow.

Measuring Distances with Strings

In order to measure any distance we need a suitable yardstick. For example, GPS uses the speed of light as a yardstick: by measuring the time a pulse of light takes to arrive from a satellite at a known location, you can obtain your distance from the satellite. Different areas of study require different yardsticks. When we talk about measuring the circumference of a compactified circle in string theory, the only suitable yardstick is a string. But there are two *different* types of yardstick we can construct from strings: we can use momentum modes or winding modes to measure distances.

In 1988, Robert Brandenberger of Brown University and Cumrun Vafa of Harvard University showed that the distances one obtains using these two measuring devices are different. Suppose a string theory is compactified on a circle of radius R. (In order to make sense of the concept of a length R, we must first define what distance means. It does not matter which string configuration we use to define distance, as long as we stick to it. So let's take the momentum mode to measure the size of the circle.) Recall that momentum modes travel freely around the circle; since we are using momentum modes, we will measure a radius R. Someone else may elect to use winding modes as a yardstick. The winding modes, though, have an energy proportional to R; according to the uncertainty principle, they can be used to probe distances of $1/R$. So, given the same compactified circle, a yardstick constructed from momentum modes will measure a large radius R; a yardstick constructed from winding modes will measure a small radius $1/R$. Both are equally valid measurements of the radius. The same argument applies to *any* distances measured using momentum and winding modes. The two yardsticks will give distances that are inversely related to each other: one large, one small.

There is a practical difference between the two yardsticks. One of the modes will be very light; the other mode will be extremely massive (so massive, indeed, that we can not hope to produce them). We necessarily measure distances with one mode — the mode that is easily produced — and we do not even recognize that a second method of measuring distances is allowed and is equally valid. This is the answer to the paradox mentioned above. It is indeed possible for galaxies to "fit" inside a universe that is an infinitesimal speck: *but you have to use the same approach when you measure the size of the universe and when you measure the size of the galaxies.* When we talk of a universe that is 10^{-96} meters wide and a galaxy that is 100 light years across, we are using winding modes to measure the first distance and momentum modes to measure the second (or vice versa). If both distances are measured using the same yardsticks, then the paradox vanishes.

First, it seems that we need to generalize the Heisenberg uncertainty principle. According to the usual formulation of the uncertainty principle, the best spatial resolution we can achieve depends on the momentum of the probe: by going to large enough momentum we can probe arbitrarily small dimensions. If the fundamental probes are strings, however, the spatial resolution one can achieve depends also on the string size. But when we put in energy at the string scale, the strings expand. We end up not probing short distances, but rather observing the propagation of large strings.

Even more interestingly, it requires that our classical notions of spacetime geometry eventually break down and must be replaced by stringy geometry. Seemingly obvious geometrical concepts, such as the difference between large and small distances, are ambiguous.

M-Theory

One must be prepared to follow up the consequences of theory, and feel that one just has to accept the consequences no matter where they lead.
Paul Dirac

Thanks to S-duality, we can understand something of the strong-coupling behavior of three of the superstring theories. As the type I string coupling constant increases, the type I theory becomes the weakly coupled heterotic SO(32) theory. As the string coupling constant of heterotic SO(32) theory increases, the theory becomes the weakly coupled type I theory. The type IIB theory is self-dual: as the string coupling constant increases, the theory can be described in terms of the weakly coupled regime. But what of the other two superstring theories? What happens in the strong-coupling limit of type IIA superstring theory and heterotic $E(8) \times E(8)$?

What Witten announced to the *Strings '95* conference, and what amazed his fellow string theorists, was his result regarding the strong-coupling limit of type IIA superstring theory. As the string coupling constant of this theory gets larger, the theory develops an extra, circular dimension! It becomes an 11-dimensional theory! Soon after this work, Witten and his colleague Petr Hořava demonstrated something similar for heterotic $E(8) \times E(8)$ theory: in the strong-coupling limit, the theory develops an eleventh dimension, but this time the extra dimension is a segment of a line.

For years, physicists had been convinced that ten was the critical dimensionality of spacetime for superstrings. Yet here was Witten, the world's premier string theorist, arguing that there is a regime in which type IIA string theory lives in 11 dimensions. So what is going on?

The first point to make is that there is no inconsistency between saying "superstrings make sense only in a ten-dimensional spacetime" and "the strongly-coupled theory lives in an 11-dimensional spacetime." There is no inconsistency because the 11-dimensional theory is not a string theory: the 11-dimensional Universe contains two-dimensional *membranes* (or *branes*) rather than one-dimensional strings. For a perturbative string theory, ten is indeed the critical number of dimensions; physicists now, though, were moving beyond perturbation theory and beyond the notion of strings as the fundamental objects.

But how could physicists fail to notice a whole new dimension in their theories? To answer that, consider the radius of the circular dimension that develops in type IIA as the string coupling constant increases. It turns out that the radius is directly proportional to the coupling constant. If the coupling constant is large (the strong-coupling limit) then the radius is large, and the eleventh dimension is manifest. But from the outset, physicists assumed that the coupling constant is small; they *had* to assume that in order to do perturbation theory, which was the only approach they understood. In fact, doing perturbation theory in the string coupling constant is tantamount to doing a perturbation expansion around a circle of radius zero. In other words, the width of the membrane was assumed to be zero right from the outset and *that* is why the extra dimension was invisible. When the width of the membrane is zero, we have a string in ten dimensions.

The second point to note is that not everyone was surprised that an 11-dimensional theory might play a role. People like Michael Duff, for example, had persevered with investigations of 11-dimensional spacetimes even when the rest of the physics community had decided that ten dimensions was *de rigueur*. The reason for a continuing interest in 11-dimensional theories was that 11 is the maximum number of dimensions in which one can write down a realistic theory containing supersymmetry. Why would nature almost use the full number of allowed dimensions, but stop one short? Furthermore, the 11-dimensional supergravity theory of Cremmer, Julia, and Scherk was for a long time the best bet for a unified theory. The supergravity approach to unification (namely, extend Einstein's theory of gravity with supersymmetry in the hope of incorporating particle physics) seemed just as valid as the string approach to unification (namely, extend particle physics in hope of incorporating gravity). As we saw earlier, it turned out that 11-dimensional supergravity theory did not quite work as a final unified theory. But physicists like Duff were sure that such a marvellous structure must play *some* role in physics. The number 11 was special.

Although Witten had found an extra dimension in type IIA string theory, what exactly was this 11-dimensional theory that was the strong-coupling limit of type IIA theory? Nine years after the *Strings '95* con-

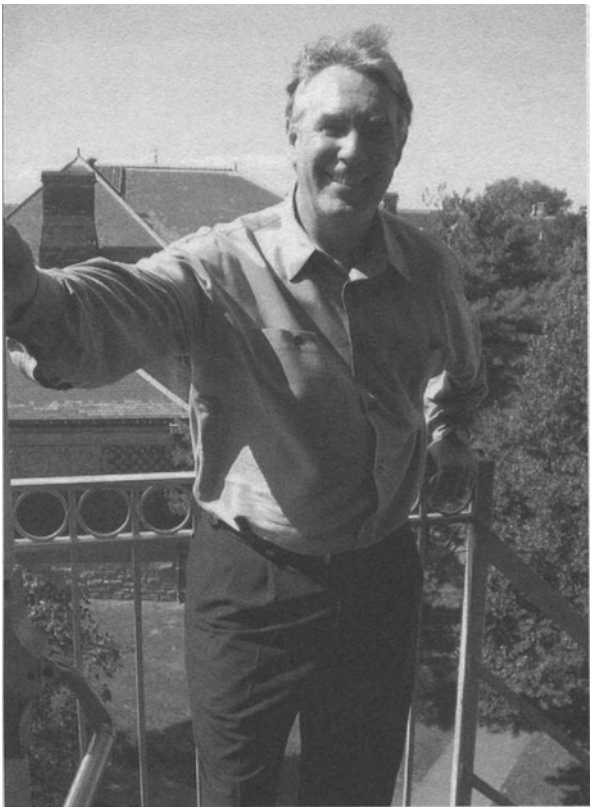


FIGURE 102 In the early 1970s, at Imperial College London, Michael Duff was a student of the Nobel prize-winning physicist Abdus Salam. Duff was a long-standing advocate of 11-dimensional theories when most of the string community seemed interested only in theories in ten dimensions. He is now the Director of the Michigan Center for Theoretical Physics.

ference, physicists still do not have an answer to this. They lack a precise definition of the theory. Using the power of dualities they have uncovered some of its properties, but a complete definition is out of reach and progress is slow. Physicists are rather like a group of archeologists who have found an earth-encrusted object; as they brush away debris they continue to learn a little more about the object, but so far the totality of the object remains hidden from view. Nevertheless, Witten was able to deduce one important fact about the theory: at low energies, the theory is described by 11-dimensional supergravity!

At this point it is worth taking stock about what has been learned.

Using a powerful type of symmetry — duality — physicists have been able to investigate the five superstring theories in previously uncharted regions. It turns out that there is a complicated web of relations between the various theories; one can start with any superstring theory and arrive at any of the other superstring theories by smoothly following the dualities. There is a sense, then, in which one can say that they are all different aspects

of the same theory. Furthermore, by considering the strong-coupling limit of two of the string theories — type IIA and heterotic E(8) × E(8) — we are driven to a completely new type of theory: an 11-dimensional theory containing membranes.

We can view this information in the following way: over the past 30 years, working in strings and in supergravity, physicists have been seeing glimpses of a fundamental, underlying structure — *M-theory*. The five superstring theories in ten dimensions, and the 11-dimensional theory discussed by Witten, are not distinct theories in themselves; rather, they are particular limits of the more fundamental M-theory. Specify a particular coupling constant and a radius for M-theory and you end up with type IIA strings, say; specify different values and you end up with type IIB strings. (See Figure 103.) So we have *one* theory, and six different limits — or vacua — of that single theory. (This is the same viewpoint as traditional superstring theory. A single theory has many thousands of different vacua corresponding to different values of the moduli fields and thus different possibilities for compactification.)

None of the six limits — the five string theories and the 11-dimensional theory — is more fundamental than the others. Duality links them all. What distinguishes these points in the hexagon is that they are limits of the underlying theory that possess two special symmetries: supersymmetry and Poincaré symmetry. What also distinguishes these limits is that they are places where physicists know how to perform calculations in M-theory. We can do M-theory calculations in these limits because there is an effective description in terms of a weakly-coupled string theory or 11-dimensional supergravity. Most of the diagram in Figure 103, though, is shaded: the shaded area represents M-theory where the coupling constant is large — the perturbative approach does not work, and the non-perturbative theory is hidden from us.

Some authors use the term "M-theory" for the top-right vertex in Figure 103 — the 11-dimensional theory whose low-energy limit is 11-dimensional supergravity. This is the theory that Witten discussed in 1995. Many more authors, though, reserve the term "M-theory" for the underlying quantum theory that subsumes the 11-dimensional theory and the five string theories. I shall follow the latter convention. Furthermore, since we are no longer dealing purely with string theory — but with a framework that contains higher-dimensional branes in addition to strings — from now on I shall use the term M-theory when discussing a unified theory. Most authors still talk of superstring theory in this context, but perhaps it is time to break that habit; whatever the unified theory turns out to be, we now know that it is more than just a theory of strings.

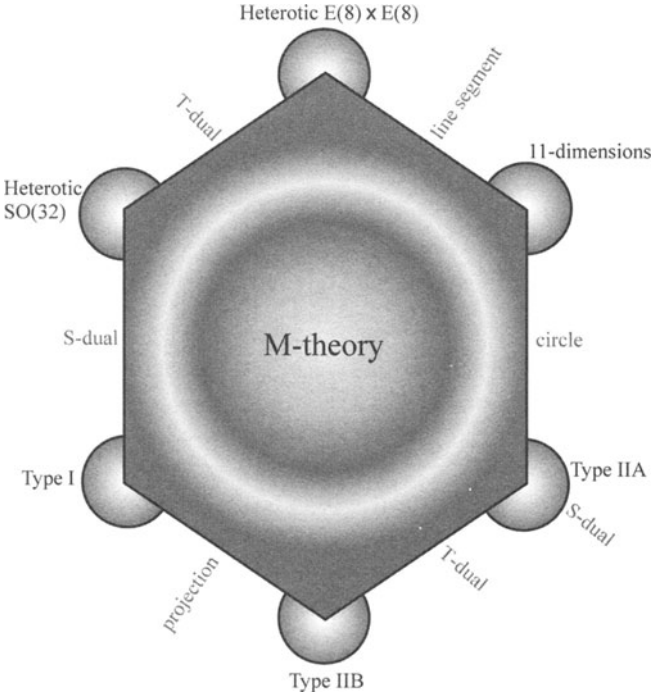


FIGURE 103 The M-theory hexagon. Each vertex is a particular limit of the underlying M-theory; there are five string theories and an 11-dimensional theory. The lines between them are labelled by the dualities that link the theories. Dualities mean we can move from one limit to any other limit. For example, if we start with type I theory (the lower-left vertex of the hexagon) we can obtain the heterotic SO(32) theory by S-duality. The heterotic SO(32) theory, when compactified on a circle, is T-dual to the heterotic $E(8) \times E(8)$ theory. The heterotic $E(8) \times E(8)$ theory is linked to the 11-dimensional theory: when the latter is compactified on a line segment, it is dual to the heterotic string. When the 11-dimensional theory is compactified on a circle, it is dual to the type IIA theory. The type IIA theory compactified on a circle is T-dual to type IIB theory. And type IIB theory, when projected in a certain way, gives us type I theory. The diagram becomes even more complex when other compactifications are considered.

CONCLUSION

Be united — united — united.
Friedrich Von Schiller, *William Tell*

I wrote earlier that the "M" in M-theory could stand for many different things: meta, magic, membrane, or a host of other words. (In 1996, four physicists — Thomas Banks and Stephen Shenker, then at Rutgers University, Willy Fischler of Texas University, and Lenny Susskind of Stanford University — formulated M-theory in terms of branes that exist in an 11-dimensional spacetime, with the spatial positions of the branes being

determined by matrices. This influential "matrix model" gave rise to another meaning for "M" — matrix.) The "M" is not pinned down because the *theory* is not yet defined. It remains mysterious. The dualities discussed above, such as S-duality and T-duality, are almost miraculous (yet another "m" word that may be appropriate for M-theory). They point strongly to the existence of a fundamental quantum theory, the various limits of which happen to be the best attempts physicists have come up with to the problem of unifying gravity with the other forces. This is a wonderful discovery, even if much of the theory remains to be uncovered. Mathematicians are equally enamored of the theory: wonderful links between different branches of mathematics have been discovered because of M-theory.

Critics of M-theory can, with some justification, point out the shortcomings of the theory. First, it is not even a theory — at least not in the sense that other parts of physics contain theories. Second, there has still been no contact with experiment. For example, we still do not know how to compute the many parameters of the Standard Model, and we still do not know why we observe a Universe with three spatial dimensions rather than some other number. The value of the cosmological constant remains an embarrassment: there is still a mismatch of a hundred orders of magnitude between calculation and experiment. Some critics decry the whole M-theory approach.

Yet such criticisms are surely harsh. There are so many factors pointing towards M-theory as being *the* unified theory that it seems reasonable at least to investigate the ideas as fully as possible. At present, those ideas can only be investigated using mathematics. But we can at least hope that those investigations will one day enable physicists to specify the theory precisely; they might then extract predictions that can lead to experimental tests of the idea. And as we shall see in the final chapter, the M-theory approach has inspired several ideas — the "wild ideas" of this book's title — that may soon be tested in experiment.

Even in the absence of experimental tests of M-theory, a crucial development in 1996 meant that M-theory came within touching distance of the rest of physics. The key to M-theory was the existence of higher-dimensional membranes, or branes. Not only were branes important in uncovering the various string-theory dualities, but M-theory was mathematically inconsistent without branes. Then in 1996, physicists showed how an understanding of branes could lead to a deeper understanding of black holes. These important developments are the subject of the next chapter.

9

OF BRANES AND BLACK HOLES

Black holes are where God divided by zero.
Steven Wright

Superstrings in ten dimensions are not by themselves the truly fundamental objects that physicists once thought them to be. Strings alone *cannot* be fundamental because two of the five superstring theories are dual to an 11-dimensional theory, yet perturbative string theories cannot be defined consistently in 11 dimensions. A truly fundamental non-perturbative theory must contain a range of basic objects in addition to strings. In fact, in order to be consistent, a fundamental theory — which we may as well call M-theory — must contain a range of objects called *p-branes*. These are non-perturbative objects, which do not correspond to the oscillations states of a string. Indeed, it was partly by examining the properties of branes that physicists were led to the notion of M-theory. The purpose of this chapter is to describe the notion of branes in a little more detail — and to describe one of the great triumphs of M-theory to date: a better understanding of black holes.

BRANES

Then, the cool kindliness of sheets, that soon
Smooth away trouble.
Rupert Brooke, *The Great Lover*

Theorists came across branes in studies of supergravity theories which, as we know, are low-energy limits of superstring theories. These supergravity theories often possessed solitonic solutions, which could be interpreted as topological defects — much as magnetic monopoles could be interpreted as topological defects in grand unified theories. No one has observed the topological defects of supergravity or grand unified theories, of course, but similar structures can be observed in a more mundane context: topological defects also appear in crystals, and defects of various kinds are most easily studied in liquid crystals. (See Figure 104.) Liquid crystals are crystals that have lost some or all of their positional order, but keep their full orientational order. Imagine shaking a rectangular box that contains a large number of toothpicks: when you open the box, the toothpicks will be facing in roughly the same direction, but will have no definite spatial organization. They are free to move, but they prefer to line up in the same direction. The rod-like organic molecules in a nematic liquid crystal behave in much the same way. By studying defects in crystals we can get some feel for the behavior of defects in supergravity theories.

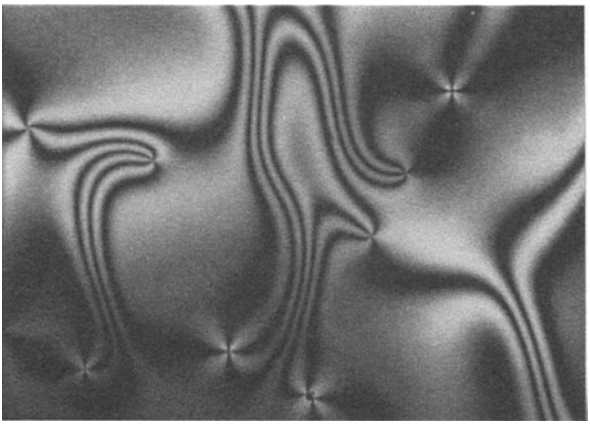


FIGURE 104 This is a photograph of a liquid crystal taken under a polarizing microscope. At several places, four dark "brushes" emerge from a point. These points are topological defects. Crystals can also exhibit line and plane defects. In string theories, higher-dimensional topological defects (branes) can occur.

The solitons, or defects, in supergravity theories were given the name generic name of p-brane. The name is a shortened form of the word membrane: the first type of brane studied was a two-dimensional sheet, or membrane, but branes with higher dimensions are possible. In fact, a p-brane is simply an object with p spatial dimensions that can move freely through spacetime.

The energy of a brane is concentrated on its p-dimensional surface, and the energy density is called the *tension* of the p-brane. Thus a 0-brane is a point particle, and in this particular case the tension is just the particle's mass. A 1-brane is a string (since strings are one-dimensional objects), and the tension is the string tension. But in addition to particles and strings, we can have 2-branes (a 2-brane is a sheet, or membrane, as mentioned above), 3-branes (which might look like a wobbly Jell-O), and branes of even higher dimension (which are difficult to describe and perhaps impossible to visualize). There is a restriction on the dimensionality of a brane: there must be room for the brane to move around in spacetime; so $p + 1$ must not exceed the number of spacetime dimensions. In our everyday world, spacetime has four dimensions; the original Kaluza–Klein model postulated a spacetime of five dimensions; superstrings live in a spacetime of ten dimensions; M-theory permits a spacetime of 11 dimensions. In each case, the maximum dimensionality of a p-brane in that spacetime must be less than the dimension of the spacetime. For example, there is no place for a 6-brane in our 4-dimensional world.

One might suppose that string theorists would have graduated to brane theories almost as soon as strings were touted as a potential tool for unification. After all, the historical progression in high-energy physics was from a consideration of zero-dimensional objects (particles) to one-dimensional objects (strings); so why not higher-dimensional objects (branes)?

In fact, the first discussion of membranes in high-energy physics predates string theory, appearing as long ago as 1962. In another of his remarkably prescient papers, Dirac considered a model in which the electron was a closed membrane, or a bubble, and it was oscillations of this membrane that generated the muon (which, as we have seen, seems to be simply a more massive version of the electron). As a model of the electron, Dirac's membrane theory did not work, but Dirac was always driven by the beauty of his equations. It turns out that the mathematics he employed to describe the behavior of the electron membrane are what modern physicists require to describe branes. So if Dirac could consider membranes back in 1962, why should not string theorists consider them in the context of unification?

Some superstring theorists did indeed investigate membranes. In 1986, a trio of physicists then based at Texas University, James Hughes, Jun Liu, and Joseph Polchinski, showed that it was possible to marry the ideas of supersymmetry with those of membranes: *supermembranes* were possible. They demonstrated that mathematics allowed the existence of one particular supermembrane: a 4-brane in a 6-dimensional spacetime. The following year, Eric Bergshoeff and Ergin Sezgin, who were then at the International Center for Theoretical Physics at Trieste, Italy, and Paul Townsend of Cambridge University, discovered the existence of a 2-brane — a closed sheet,

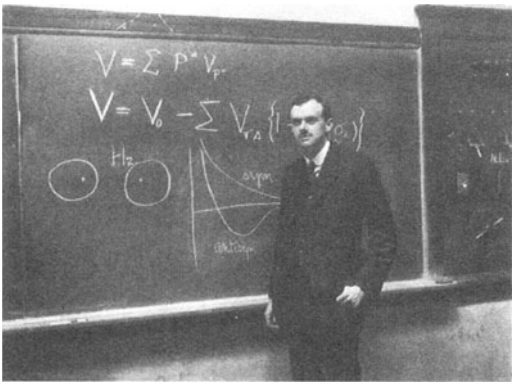


FIGURE 105 Paul Dirac considered the theories of membranes in high-energy physics as long ago as 1962. Although the specific model he investigated was unsuccessful, as always Dirac was motivated by the beauty of his equations; some of the mathematics he developed to describe membranes are used by physicists who are now investigating M-theory.

like a bubble — moving in an 11-dimensional spacetime. Indeed, it quickly became apparent that many types of brane were possible, but the demands of supersymmetry placed severe constraints on the allowed dimensionality of the brane. Physicists were soon able to complete a *brane scan*: a complete plot of the allowed dimensions of a brane in any particular spacetime dimension. (See Figure 106.) We see that branes can exist in any dimension up to and including 11 (although whether they actually exist in a particular theory depends upon the details of that theory).

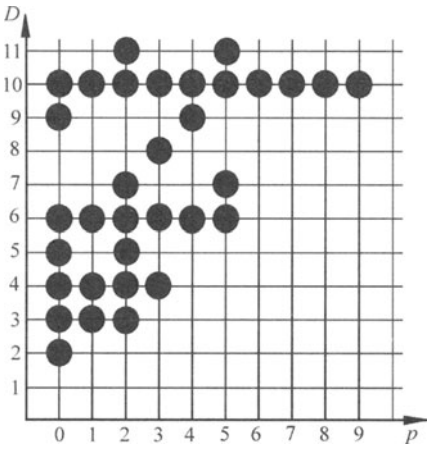


FIGURE 106 Supersymmetry places restrictions on the dimension of a brane (p) and the dimension of the spacetime (D) in which it lives. The points on this grid show the allowed types of p-brane in spacetimes with up to 11 dimensions.

The brane scan provided a reason for ignoring membranes: it allows for extended objects (a 2-brane and a 5-brane) in a spacetime of 11 dimensions. But string theorists had already decided on the critical dimensionality of spacetime: it was 10, not 11. Why bother with membranes if they can live in a spacetime that superstring theory says cannot exist? In essence, string theorists were saying that the 1-brane in 10 dimensions was the only point on the brane scan that made physical sense.

There were other problems with branes. One of the key points of string theory is that it contains a massless spin-2 particle — the graviton. According to early work on extended objects, however, it was impossible to construct a theory based on branes that contained a massless spin-2 particle. Branes, in other words, could not describe gravity. Furthermore, Green and Schwarz had shown that certain superstring theories in ten dimensions could be free from anomalies. So why bother with branes?

In 1987 Sezgin and two colleagues — Itzhak Bars of the University of Southern California and Christopher Pope, then at CERN — demonstrated that a supersymmetric theory based on branes *could* contain a massless spin-2 graviton, but only if the brane existed in 11 dimensions. There was that number 11 cropping up again. Bars and Pope then went on to show that *all* supersymmetric theories contain unavoidable anomalies, except for string theories in 10 dimensions and brane theories in 11 dimensions. Yet again, 11 dimensions were looking as important as 10 dimensions. Another stunning result came in 1987. A group of physicists collaborating while at CERN — Michael Duff, Paul Howe, Takeo Inami, and Kellogg Stelle — considered the 2-brane moving in an 11-dimensional spacetime. They showed that, if 1 of the 11 dimensions was a circle, then one could wrap the brane around it. If the radius of the circle was made sufficiently small, the brane looked like a string in 10 dimensions! More precisely, upon compactification of the 11-dimensional theory one had type IIA superstring theory!

The string community took real notice of branes when it became clear that p-branes can be thought of as topological solitons, appearing as solutions to supergravity theories. Branes are stable and inherently nonperturbative. In 1990, for example, Andrew Strominger — one of the four physicists who first examined the Calabi–Yau compactification of the heterotic string — showed that the equations of the heterotic string can yield a 5-brane as a soliton solution. He also showed that as the mass of the soliton increased, the size of the string coupling decreased. This suggested a duality relation: a strongly coupled heterotic string was dual to a weakly coupled 5-brane. Again, this was a conjectured duality, but Duff and others, using different arguments, proved the conjecture beyond a reasonable doubt. As we saw in Chapter 8, physicists learned in the 1990s how to establish the validity of various duality conjectures, and the properties of branes that were BPS states were key to this. It was the study of branes that led physicists beyond mere strings. Strings were important, certainly, but they were not the whole story. It became clear that we should not be studying a theory of just 1-branes in 10 dimensions, or even 2-branes in 11 dimensions. There was a place for 3-branes in 6 dimensions, 4-branes in 9 dimensions, and all the other combinations allowed by the brane scan.

D-BRANES

They must have the defects of their qualities.
Honoré de Balzac, *Le Lys dans la Vallée*

In 1989, at Texas University, Joseph Polchinski and his students Robert Leigh and Jin Dai discovered a new type of brane. They were looking at topological defects: extended solutions to string theory equations that are stable by virtue of topology. At the time they did not realize the importance of their discovery, but in 1995 Polchinski used dualities to reexamine the properties of that brane. It was only then that he realized the significance of what his group had discovered.

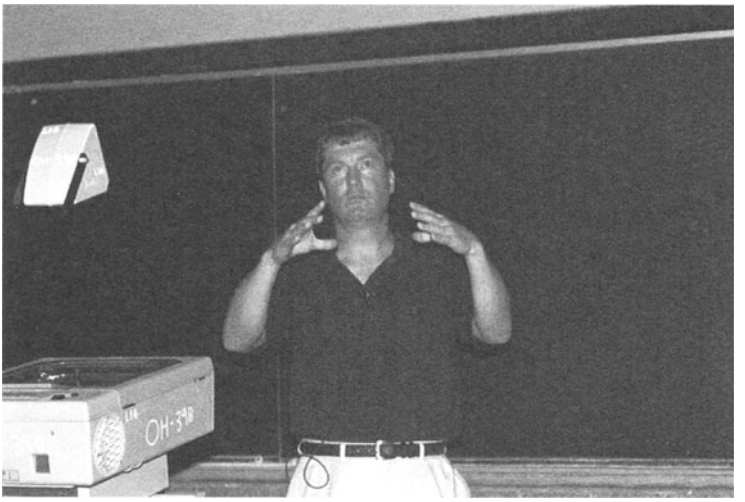


FIGURE 107 Joseph Polchinski's discovery of D-branes led to a simple and precise description of many non-perturbative phenomena in string theory.

Polchinksi and his students had been studying the boundary conditions obeyed by superstrings. Boundary conditions — what happens at the end points of a string — are clearly important whenever one tries to describe string behavior. A guitar string, for example, must obey the boundary condition that its two ends are fixed. In type II superstring theories, the closed strings obey periodic boundary conditions, because the strings loop back onto themselves. An open superstring of type I theory can obey one of two boundary conditions. Either the ends are free to move about in spacetime or else the end points of the string are fixed on a particular manifold. (The two end points need not end on the same manifold, although they can.) If the string obeys the latter condition then it is said to obey the *Dirichlet*

boundary conditions, which are named after a nineteenth century mathematician who investigated problems related to the boundaries of objects. The end points of the string can move, but only on the manifolds to which they are attached. The manifold is called a D-brane (D for Dirichlet).

FIGURE 108 Johann Dirichlet was a nineteenth century mathematician who investigated problems associated with boundary conditions.

A D-brane is simply an object upon which an open string can end. D-branes are dynamical objects, which can bend and move around. (In particular, they interact with gravity: they couple to closed strings which, as we saw in Chapter 7, are responsible for gravity. When a closed string graviton interacts with a D-brane, the string opens up and during the interaction it becomes an open string with end points that move on the brane.) D-branes come in various shapes and sizes: they can be point-like; they can be extended and have a finite size in some directions; or they can be infinitely extended in some directions.

If the dimension of the D-brane is important, then we can call it a Dp-brane (p is the dimension). In superstring theory, which is consistent only in a spacetime of ten dimensions, one can have D-branes with a dimension ranging from -1 to 9. Figure 109 shows a D0-brane (which is a point), a D1-brane (which is a curve), and a D2-brane (which is a sheet).

D-Branes in String Theories

For the record, the D-branes in the five perturbative string theories are as follows. Type I string theory has D1, D5, and D9 branes. Type IIA string theory has D0, D2, D4, D6, and D8 branes, while type IIB string theory has D-1, D1, D3, D5, D7, and D9 branes; type IIA and IIB theories are related by T-duality. Neither of the heterotic string theories contain D-branes.

D-branes were found to have many interesting properties (which, at least weak coupling, could be uncovered in much the same way that the properties of strings were uncovered). As we have already mentioned, a D-brane is a dynamical object: it can move and it can interact with other branes. A D-brane has mass and possesses various charges (like electric or magnetic

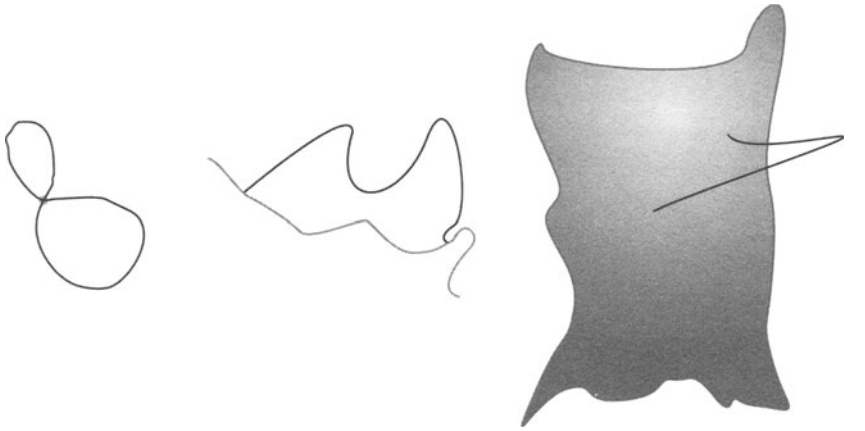


FIGURE 109 The leftmost figure shows a D0-brane (the tiny gray dot) with two strings (the black curves) attached to it; the strings are stuck to this brane. The middle figure shows a D1-brane (the gray curve) with one string attached to it. Again, the end points of the string are confined to the brane, although they can move along it. The rightmost figure shows a D2-brane with an attached string. The end points of the string are restricted to moving along the two dimensions of the brane.

charges). Since open strings attach to a D-brane, the effective theories on a brane are gauge theories — the type of theory we know describes our large, low-energy, everyday world. Crucially, D-branes are BPS states; we have seen how BPS states allow physicists to gain insights into many non-perturbative properties (and Polchinski's group took full advantage of this to learn more about D-branes). One further important property of D-branes is that, if two branes are parallel to one another, the gravitational attraction between them is exactly cancelled by the repulsion due to their charges. There is no net force between parallel D-branes. You can imagine stacking them together like a deck of playing cards. As we shall soon see, this behavior can be put to interesting use.

Most important of all, physicists understood enough of the theory to let them study how branes interact in a variety of situations. This led them to perhaps the best piece of evidence that, with M-theory, they were on the right track: with D-branes they finally had an inkling about the quantum mechanical properties of black holes.

BLACK HOLES

Black's not so black.
George Canning, *New Morality*

The gravitational force is extremely weak: far, far weaker than electromagnetism, which is the only other long-range force in nature. If you want a demonstration of just how weak gravity is, rub a balloon on your clothing for a few seconds so that it picks up a static electric charge and then use it to levitate a piece of tissue paper. Think about it: the electrical attraction on the paper which is generated from a few particles in the balloon can overcome the gravitational attraction on the paper from the entire Earth. And yet, despite its weakness, it is gravity rather than electromagnetism that dominates the dynamics of the Universe. The reason is that the electromagnetic charge comes in two opposite forms — positive and negative — and in large-scale collections of matter the two charges appear in equal number; thus matter on a large scale is usually electrically neutral. The gravitational "charge" — mass — comes in only one form, and the gravitational force is always attractive. Add more matter to an object and its mass will increase whereas its overall electrical charge will tend to zero. And since the gravitational force generated by an object depends upon its mass, gravitational fields can become extremely large. As a result we feel no electrical force from the Moon, but the Moon's gravitational force causes tides to rise in Earth's oceans.

According to general relativity, there is nothing to stop you from piling mass upon mass so that the gravitational force on an object rises without limit. Soon after Einstein completed his theory of general relativity, the German astronomer Karl Schwarzschild, who was by then on the Russian front serving as an artillery lieutenant in the German army, showed that if enough mass is crammed into a small enough volume then spacetime becomes so warped that *nothing*, not even light, can escape from the grip of gravity. These warped spacetime regions are, of course, *black holes*.

Schwarzschild — and many astronomers after him — thought that his theoretical solution was physically meaningless; many astronomers did not believe in the physical reality of black holes. However, there is now overwhelming evidence that black holes exist in nature. Astrophysicists have detected supermassive black holes at the center of many galaxies, and our own galaxy probably possesses a black hole at its core. (You cannot actually see a black hole, but you can see the radiation emitted by material just outside a black hole as it spirals towards its destruction. In fact, it is by measuring the rotation rate of the spiralling material that you can estimate the central mass of the object. And by measuring the radius of the spiralling disk, you can estimate the radius of the central object. These two pieces of

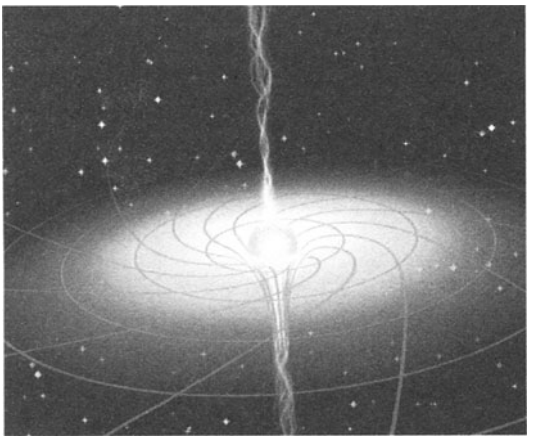


FIGURE 110 An artist's impression of material spiralling into a black hole. Once inside the event horizon the material cannot escape. But at the poles of a black hole, jets of matter can escape from a region just outside the horizon. Black holes are the inevitable final state of massive stars; if, at the end of its life, a stellar remnant is more than about three times the mass of the Sun, it will undergo gravitational collapse and end up as a black hole. Supermassive black holes, with a mass millions of times greater than our Sun, are commonly found in the cores of galaxies.

information reveal the density of the central object and tell astrophysicists they are dealing with black holes. The mass of the central object is always so large and the volume is always so small that the equations of general relativity tell us that these objects *must* be black holes. They can be nothing else.)

We have nothing to fear from astrophysical black holes. Consider a solar-mass black hole, for example. At a distance of 93 million miles — the mean distance of the Earth from the Sun — the black hole would curve spacetime in exactly the same way the Sun curves spacetime. Replace the Sun with a solar-mass black hole, and the planets would orbit in exactly the same way they do at present, although the solar system would be a much colder place. It is only when one gets close to the black hole that spacetime becomes *really* curved. And it is only when one passes the event horizon — the boundary in space at which the velocity required to escape from the gravitational clutches of the black hole becomes equal to the speed of light — that all hope of escape from the hole is lost.

The real threat posed by black holes is to the theory of general relativity. The threat comes from three sources.

First, general relativity says that a black hole contains a *singularity*, a point at which the laws of physics break down. A singularity is usually cloaked by an event horizon. Stay on the right side of the event horizon and you are safe, but stray to the other side and you will eventually hit the singularity. It seems there are places in the Universe where we cannot hope to apply our physics, and that is disturbing.

Second, although astrophysical black holes are massive — anywhere from a few solar masses up to several million solar masses — there is nothing in Einstein's theory to forbid the creation of black holes with a much

smaller mass. If you can find some way to crush *any* mass into a small enough volume, then inevitably you will create a black hole. With astrophysical black holes we can happily use general relativity to understand and describe their properties: the holes are so large that quantum effects are irrelevant. But for small black holes, quantum effects become increasingly important. For a black hole with a Planck mass, which is about 10^{19} times the proton mass, we are talking about an inherently quantum object. So even if we ignore the problem of the singularity, there are objects — Planck mass black holes — that *must* be described by a quantum theory of gravity. We must give up general relativity and use a more appropriate theory; M-theory, perhaps?

Third, there is a problem with black hole entropy. This problem is a little more subtle that the other two and deserves some further description.

Black hole entropy

The idea of black hole entropy is rather esoteric. The notion of entropy itself, however, springs from the most mundane of observations. The first glimmerings of the notion began in the early days of *thermodynamics* — originally, the study of the movement of heat. The concept of entropy was clarified as scientists began to study steam engines. Engines were one of the most important areas of study at the time. The more one understood the science behind steam engines, the more efficient one might be able to make them. That was the hope, anyway.

As far back as 1824, Sadi Carnot, a young French scientist, analyzed the workings of a steam engine in detail. He tried to calculate the amount of work one could obtain from any given engine. Carnot realized that the temperature difference between the hot and cold regions in an engine was a measure of the available energy. But in any closed system, heat flows from a hot region to a colder region, and thus the temperature difference decreases with time. The available energy, therefore, decreases with time. Since the total energy of a closed system is constant, the *unavailable* energy always *increases* as time passes. Carnot's observations were specific to heat engines. In 1850, though, the German physicist Rudolf Clausius generalized Carnot's findings. Clausius called the unavailable energy of a system the *entropy* of that system. According to Clausius, the entropy of *any* closed system always increases with time. This statement — which is a version of the famous second law of thermodynamics — holds for all closed systems, from steam engines to refrigerators to planet Earth. The concept of entropy is one of the deepest in all of physics.

But what exactly *is* entropy? In the 1890s, the Austrian physicist Ludwig Boltzmann deduced the second law of thermodynamics from more

basic, fundamental laws. Boltzmann argued that, if you accept that a gas is composed of molecules randomly bouncing around, then you can use statistical techniques to calculate various quantities of interest about the gas — including its entropy. In Boltzmann's picture, the entropy of a system could be thought of as a measure of the disorder of that system: the higher the entropy, the more disordered or random it must be. So a fresh deck of playing cards has low entropy because it is extremely ordered; after a few shuffles, the deck has high entropy because it is extremely disordered. We can also think of entropy as the number of different ways the energy of a system may be generated from the combined energy of its constituent elements. So a gas close to absolute zero has a very low entropy, because there are very few different configurations of the gas molecules that can contribute to the energy of the system as a whole. If the gas molecules are all frozen in place, unable to move, then we have a highly ordered system. Raise the temperature, though, and the gas molecules begin to zip around. Many different states of the system can give rise to the energy, and the entropy rises.

Boltzmann's idea came long before the notion of molecules was firmly accepted by the scientific community. Proof of the molecular nature of matter did not come until around the time of Boltzmann's death, with phenomena such as Brownian motion — the random jiggling of pollen grains suspended in water, which can be seen by using a microscope — being explicable only within a molecular framework. But even before molecules were directly observed, the power of Boltzmann's insight made the existence of molecules seem more likely. His idea gave a simple physical explanation for the meaning of entropy and, moreover, gave a precise mathematical recipe for calculating the entropy of a system: simply use well-understood statistical techniques to count all the microscopic configurations that the constituents can adopt.

Boltzmann's work required very little modification to be used in quantum mechanics. The "microscopic configurations" of a system are called, in the jargon, *microstates*. So to calculate the entropy of a quantum system you simply count the number of quantum microstates. In fact, in appropriate units, the entropy of system is just the logarithm of the number of microstates. Entropy is thus a precisely defined quantum mechanical concept. And to repeat: entropy is one of the most useful and general concepts in physics.

Let's now return to the matter in hand: black holes. In the 1960s, Stephen Hawking showed that the area of the event horizon of a black hole can never decrease; in any interaction, it can only get larger. In 1972, Jacob Bekenstein, who was then at Princeton University, pointed out that this

behavior is analogous to the entropy of a system according to the second law of thermodynamics. In fact, he proposed that the area of the event horizon gives a precise measure of the entropy of a black hole.

The Area of an Event Horizon Always Increases

Suppose a black hole of mass M_1 and horizon area A_1 coalesces with a black hole of mass M_2 and horizon area A_2. The mass of the new black hole is simply $M_1 + M_2$: the masses of the black holes add. But the event horizon area is proportional to the square of the mass. So the new area A is proportional to $(M_1 + M_2)^2$ — which is *greater* than $A_1 + A_2$. The area of the black hole increases. If a star falls into a black hole, the area of the black hole increases; if a planet or a comet or a grain of dust falls into a black hole, then the area increases. If *anything* falls into a black hole, the area increases.

For Bekenstein, the idea that a black hole has entropy seemed eminently reasonable. Suppose you pour the contents of a thoroughly disordered system (a system with high entropy, in other words) into a black hole. You are left holding an empty system, which is certainly ordered and thus possesses low entropy. Where has the entropy gone? If entropy can only increase, it must have gone *somewhere*. Well, according to Bekenstein, the area of the black hole will have increased — and *this* is where the entropy has gone.

Hawking initially dismissed Bekenstein's idea. There were three problems with it. First, a black hole seemed to be too simple an object to allow for large entropy: once you have specified its mass, its charges, and its spin, you have said pretty much all there is to say about it. (Interestingly, this is similar to how we described the fundamental particles in an earlier chapter. Once you have specified a few basic properties, you have said all there is to say about a fundamental particle. Perhaps black holes and fundamental particles are linked?) A system with high entropy is a highly disordered system, but a black hole is too simple an object to possess much disorder. Second, we know that the entropy of a system is essentially the number of its quantum microstates; at least, it is for every physical system we know of, so presumably it is the same for black holes. But black holes are the quintessential objects of general relativity, and neither Hawking nor anyone else knew how to marry quantum mechanics with general relativity. Third, according to thermodynamics, if a black hole has an entropy given by the area of its event horizon then one must associate a temperature with

the black hole — a temperature that is proportional to the strength of the gravitational field at the event horizon. But if a black hole has a temperature then it must radiate energy. That would seem to be the killer blow for Bekenstein's idea: black holes don't radiate, they *absorb*. Black holes are black, aren't they?

FIGURE 111 In 1974, Stephen Hawking of Cambridge University announced that black holes are not really black: they radiate energy. A deeper understanding of Hawking's result is slowly becoming available thanks to M-theory.

In 1974, Hawking made perhaps his greatest contribution to physics. He showed that black holes are not black at all. Hawking looked at what happens close to the event horizon of a black hole, but he did so from the point of view of quantum mechanics. In a certain sense, he was able to combine quantum theory with gravity; it was not a full quantum theory of gravity by any means, but it *was* a quantum mechanical description of what happens in the strongly curved spacetime near the event horizon of a black hole. So what does happen? Well, we know that quantum mechanics allows the creation of a virtual particle–antiparticle pair. As long as the pair annihilate in a timescale given by the uncertainty principle, nature does not care about their fraudulent existence. Suppose, then, that a pair of virtual photons pops into existence. (Remember that the photon serves as its own antiparticle.) If the photons are created just outside the event horizon, then the gravitational pull of the hole can suck one of the photons over the horizon. Once the photon is on the other side it can no longer communicate with the rest of the Universe; the photon with which it was created no longer has an annihilation partner. The photon that remains on the safe side of the horizon becomes real; it gets an input of energy from the gravitational field, and it then travels away from the hole. To an observer looking at the hole from afar, the hole appears to have emitted a photon. The combined effect of a ceaseless process of particle pair creation,

followed by the destruction of one particle and the emission of the other, means that black holes radiate. They glow. They have a temperature.

Hawking's calculations also showed that the more mass a black hole possesses, the lower its temperature. Indeed, the temperature of any astrophysical black hole is only a tiny fraction of a degree above absolute zero. We cannot detect such radiation, and so Hawking's discovery has no practical implications. (The smallest, and therefore the hottest, black holes to be found in astronomy are likely to be those formed in the aftermath of a supernova. Such black holes have a temperature of about 10^{-7} kelvins.) But the importance from a theoretical viewpoint is that black holes must have a temperature — and therefore they must have an entropy. In fact, in appropriate units, the entropy of a black hole is just one quarter of its horizon area. The entropy of a black hole, even the smallest astrophysical black hole, is enormous. For an astrophysical black hole with a mass of 10^{30} kilograms, the number of microstates must be of the order $10^{10^{16}}$, which is an unimaginably large number. Black holes are *highly* disordered objects, and Bekenstein's idea is thus perfectly acceptable.

But if black holes are disordered, what are the microstates that allow this disorder? General relativity says that black holes have a mass, charge, and spin. What else could there be to allow for disorder — for entropy — in a black hole? It was 23 years after Hawking's discovery that physicists could at last begin to answer this question.

In 1996, Andrew Strominger and Cumrun Vafa, string physicists we have met before in our story, realized that they had a tool with which they could develop a microscopic description — and thus calculate an entropy — of a black hole. The appropriate tool had to be M-theory, since it was a quantum theory that included gravity. (The reason no one had previously attempted to calculate black-hole entropy using traditional superstring theory — which is also a quantum theory that includes gravity — was that there was no way of using it to compute answers in a strongly coupled regime. And a microscopic black hole is an object that is certainly in the strong coupling regime.) Strominger and Vafa knew how to use D-branes to investigate non-perturbative regimes. D-branes were the key to unlocking the microscopic properties of black holes.

As we noted earlier, one of the properties of D-branes is that their gravitational attraction and electromagnetic repulsion cancel each other. This means that you can stack D-branes one on top of another and make an object that is as massive as you wish. (These massive objects are a product of the theoretician's ingenuity and imagination; they are not something that we can ask astronomers to go and search for.) For example, you could wrap

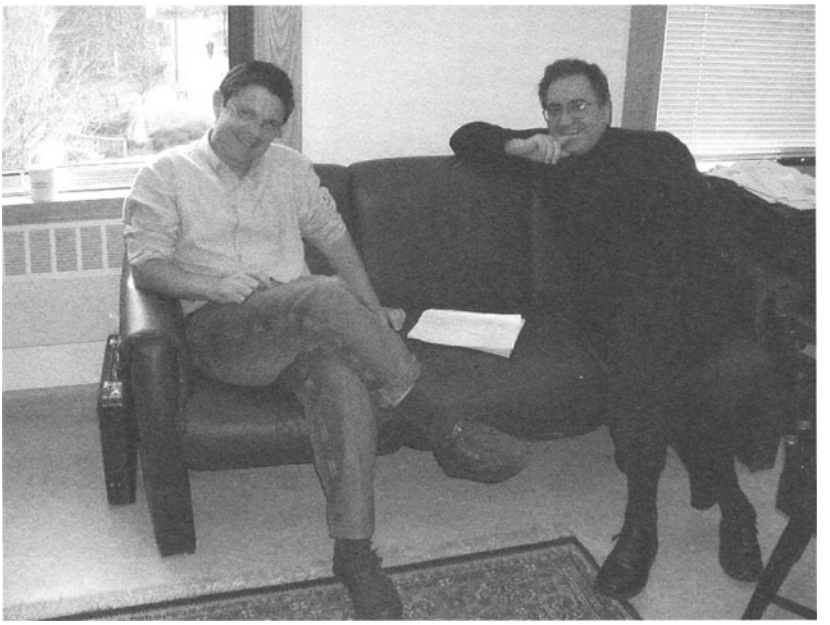


FIGURE 112 In 1985, Andy Strominger (left) was a coauthor of a paper along with Candelas, Horowitz, and Witten that was one of the high points of the First Superstring Revolution. His 1996 paper with Cumrun Vafa (right) has been one of the high points of the Second Revolution.

D1-branes — strings — around a compactified dimension; keep adding on D1-branes in this way and you end up with something that is extremely massive. Or you could wrap a higher-dimensional D-brane around a multidimensional compactified space; keep adding the D-branes and again you end up with an object that is extremely massive. The point is that, if you wrap a D-brane around the compactified dimensions in just the right way, and if you keep on doing it, then you end up with an object that behaves just like a black hole. To repeat: such a D-brane construction is *not* an astrophysical black hole; it is not something that astrophysicists can find in the remnants of supernova explosions or in the cores of violent galaxies. Rather, it is a gedanken black hole. But we have seen the power of gedanken experiments throughout this book; in the case of the Strominger–Vafa gedanken black hole, we have an object that can be investigated with the full armory of the theoretician's arsenal.

Many different types of black hole are allowed in string theory — holes with different dimensions, holes with different charges, and so on. Strominger and Vafa first had to find one with properties simple enough for them to work with. They looked, therefore, at the simplest type of gedanken black hole: the *extremal* black hole. An extremal black hole pos-

sesses a charge that cancels the evaporation process of Hawking radiation. Whereas interesting black holes are "gray" and emit radiation, extremal black holes have zero Hawking temperature and just sit there. They also have the minimum allowed mass for the charge they possess, which makes them a type of BPS state. These properties make extremal black holes rather boring objects, but they were exactly the sort of object that Strominger and Vafa knew they could work with.

Strominger and Vafa then had to search for a means of constructing an extremal black hole for which they could actually do the required computations. They eventually found they could give a complete description of a five-dimensional black hole. They started out with the familiar ten dimensions of type IIB string theory and compactified five of the dimensions on a torus (a doughnut). They then wrapped D5-branes and D1-branes around the compactified dimensions, creating a bound state of twisted branes; as they kept adding branes, the mass of the object kept rising. They created a black hole: not an astrophysical black hole from astronomy, but a gedanken black hole from imagination. The beauty of the gedanken black hole, of course, was that its creators knew everything about it. They had built it, after all. Strominger and Vafa could thus use standard statistical techniques to count the number of microstates of this tangle of branes — they could calculate the entropy.

The calculation of black hole entropy by counting microstates was impressive — akin to Boltzmann's calculation of the entropy of a gas by counting configurations of molecules — but it would mean little if it gave the wrong answer. Interestingly, Strominger and Vafa did not know the Bekenstein–Hawking entropy of their black hole; no one did. No one had thought to calculate the area of the event horizon of such a hypothetical extremal five-dimensional black hole. When Strominger and Vafa computed the Bekenstein–Hawking entropy they found that the number agreed *exactly* with their previous brane-based calculation.

This was an amazing result!

Strominger and Vafa had two descriptions of black hole entropy, and the two descriptions agreed perfectly. The Bekenstein–Hawking expression for the entropy might be interesting, but it provided no deep understanding of black holes, just as the Clausius expression for the entropy of a gas provided no understanding of the nature of a gas. The Strominger–Vafa expression promised to provide an understanding of black holes in terms of deeper concepts, just as Boltzmann's expression for the entropy of a gas provided a deeper understanding of gas in terms of molecules.

There was always the chance that this result was a coincidence. To confirm the significance of the approach, physicists needed to consider several different types of black hole, since an extremal black hole in five dimen-

sions is about the most boring object imaginable. For example, what about black holes that emit Hawking radiation? Does the approach of counting microstates work in this case? After considering the simplest possible black hole, Strominger and his colleague Gary Horowitz, his coauthor of the famous 1984 paper that studied string theory compactification on Calabi–Yau manifolds, constructed the next most complicated type of black hole: a slightly non-extremal hole in five dimensions. This would have a Hawking temperature and would therefore radiate energy. Independently, the Princeton physicist Curtis Callan and his student Juan Maldacena studied the same problem. Both groups calculated the radiation rate from their tangle of D5-branes and D1-branes and found that it agreed exactly with the expression for Hawking radiation from a general relativistic black hole. At about the same time, it was shown that an extremal five-dimensional *rotating* black hole had the correct Bekenstein–Hawking entropy. Two approaches giving identical results; this was surely not coincidence.

Physicists soon created four-dimensional black holes, which were slightly closer to the astrophysical black holes of the real world (though these creations were still not the beasts found in astrophysics). The calculations for Hawking temperature and radiation of four-dimensional black holes agreed well with the general relativistic results. And when Samir Mathur, of the Tata Institute in Mumbai, and Sumit Das, then at MIT, calculated the rate at which a warm black hole would cool to absolute zero, they found perfect agreement between the M-theory approach and general relativity.

Then Strominger and Maldacena got together to calculate the total energy spectrum of black hole radiation. When you do the calculation in general relativity, in the manner of Hawking, you find that at its horizon a black hole emits a thermal spectrum — exactly the spectrum that a black body radiates due to its temperature. Far away from the black hole, however, the spectrum is altered because the emitted particles must climb out of the hole's gravitational well. The corrections to the spectrum are known as gray body factors, and they depend upon frequency in a rather complicated fashion. Strominger and Maldacena calculated the frequency-dependent corrections using both the classical Hawking method and an M-theory method. The result? Once again, there was perfect agreement. This result was really quite remarkable, since the gray body factors are tortuous and convoluted functions of frequency; you simply do not get this sort of result without there being some deep connection between theories.

There seems, then, to be a deep correspondence between the surface of a black hole and M-theory. But what could this correspondence be telling us? The answer to that marks the high point of M-theory to date.

10

OUR WORLD AS A HOLOGRAM

What jobs are there in a backwards reality for a dead hologram?
Arnold Rimmer, *Red Dwarf*

Gerardus 't Hooft is one of the deepest and most original thinkers in modern physics. As we saw in earlier chapters, his work on the renormalization properties of gauge theories helped establish the Standard Model of particle physics. His work on monopoles led to a deeper understanding of grand unified theories. And he has made many other profound contributions that the physics community continue to explore (but which cannot be explored here for lack of space). For about ten years after 1984, however, 't Hooft was out of the mainstream of theoretical research in high-energy physics. The discovery of anomaly cancellation by Green and Schwarz led most theorists to investigate superstrings, but 't Hooft, in his own words, could not quite fathom the fundamental logical coherence of superstring theory. He believed superstring theory presented as many problems as it solved.

So 't Hooft decided to try a different route to the problem of finding a quantum theory of gravity. This led him to one of the wildest, most interesting ideas in recent physics: the holographic principle. Leonard Susskind, one of the founders of string theory, immediately elaborated upon 't Hooft's idea. But it was not until the rest of the high-energy physics community realized that M-theory seemed to obey the holographic principle, that 't Hooft was back in the mainstream.

Mainstream or not, the idea that our Universe may obey the holographic principle leads to some truly bizarre possibilities. In some formulations it could mean that you, gentle reader, are a hologram; the three-dimensional world that we see around us could be an illusion; those who believed that the Earth is flat would, in a way they are unlikely to have been able to comprehend, be correct.

The starting point for these ideas is information — and its intriguing relationship to entropy.

ENTROPY

Maximum disorder was our equilibrium.
T. E. Lawrence, *Seven Pillars of Wisdom*

We have already looked at the concept of entropy. It is a down-to-earth concept if we apply it, as Clausius did, to objects like steam engines. As we stressed, this type of entropy — *thermodynamic entropy* — is one of the most useful concepts in all of science. According to Boltzmann, however, we can understand entropy at a deeper level than just as a measure of the unavailable energy in a particular system. Entropy is the number of distinct arrangements of microstates in a system that give rise to the same energy of that system. The entropy of a volume of gas, for example, is the number of different arrangements of molecules that gives rise to the same energy for the gas. The units of Boltzmann's entropy, like the thermodynamic entropy of Clausius, are energy divided by temperature.

There is another definition of entropy. It was developed by Claude Shannon, a researcher at Bell Laboratories, who introduced the concept of entropy to a branch of applied mathematics called information theory in 1948. Shannon was interested in ways of quantifying the amount of information carried by a message. Having a measure of the *amount* of information is not necessarily the same as the possessing a measure of the *value* of information; the words "I love you" constitute a string of just ten ASCII characters, but in the right context they are priceless. To put it less romantically, the value of information is highly context-dependent. Nev-

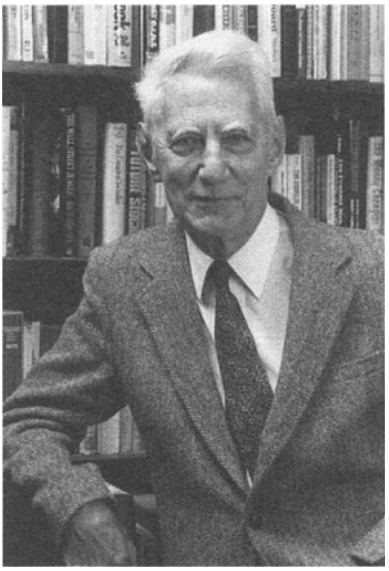


FIGURE 113 Claude Shannon laid the framework for modern telecommunications in his 1948 paper, "A mathematical theory of communication." Shannon's view of information has increasingly influenced the thinking of physicists.

ertheless, it is clear that an objective measure of the amount of information possessed by a given message would be extremely useful.

Shannon showed that the amount of useful information carried by a message is equal the number of answers to yes/no questions — the number of *bits* — needed to encode it. (For example, suppose a lottery has been entered by eight of your friends — Abigail, Bill, Catherine, David, Emily, Fred, Gillian, and Harold. Your spouse drew the lottery and knows who holds the winning ticket. You need ask your spouse only three yes/no questions to discover the identity of the winner. You could, for instance, ask whether the winner is female. A *no* answer cuts the possibilities to 4. Then ask whether the winner's first name has four letters. A *yes* answer cuts the possibilities to 2. A third question — is the winner the taller of the two — gives the answer. So the information content of the message announcing the lottery winner is 3 bits. This particular example is rather artificial, of course, but the principle works for any example.) Shannon's definition of information now plays a key role in the telecommunications industry and in cryptography, and it also finds a place in linguistics and phonetics. Increasingly, theorists are arguing that this type of information also plays a fundamental role in physics.

The equation describing the useful information carried by a message has the same form as Boltzmann's formula for thermodynamic entropy so Shannon, after advice from the famous mathematician John von Neumann, felt free to call his quantity entropy. We can thus talk of *Shannon entropy*, the unit of which is bits — the same bits that are used in computing.

It is often a mistake to recycle terminology like this. After all, the number of answers to yes/no questions has seemingly nothing to do with the flow of heat, so why should thermodynamic entropy and Shannon entropy have anything at all in common? Perhaps we should not use the same word to describe two such very different concepts? Upon further reflection, however, we can see that Shannon's entropy of information *does* have a connection to Boltzmann's entropy of thermodynamics: we require a certain amount of Shannon information in order to implement a particular arrangement of microstates that would be counted in Boltzmann's formula. So maybe these two concepts *do* deserve the same name? Maybe they are, at a deep level, the same thing?

There are two immediate obstacles standing in the way of treating Shannon entropy and Boltzmann entropy on an equal footing. The first obstacle is that the two entropies are measured using different units: bits for Shannon entropy, energy divided by temperature for Boltzmann entropy. Fortunately, this doesn't turn out to be an issue because, with appropriate conventions, one can use the same units for the two entropies. The second obstacle is the different scales used for the two concepts: Shannon entropy is typically much smaller than Boltzmann entropy. For example, the words in the book you are holding possess a Shannon entropy of about 10^7 bits; assuming that you are reading at room temperature, the Boltzmann entropy of the book is roughly 10^{23} bits. This is a *vast* difference and would seem to preclude any possibility of equating the two entropies.

The source of this difference in scales is illuminating, and immediately suggests a way forward. The two entropies count different sorts of things, different *degrees of freedom*. The Shannon entropy of this book is concerned only with the amount of information carried by the book; the electronic manuscript that I supply to the publishers can be represented by a series of bits — 0s and 1s. The Boltzmann entropy of the book, on the other hand, depends upon the state of motion of the trillions of atoms that constitute the fabric of the book; the degrees of freedom here are quite different to the binary, on/off, degrees of freedom used in calculating Shannon entropy. It is little wonder that the scales of the entropies are so different.

A book is a fairly bulky method of storing information, however. We can store the same amount of information in a small piece of computer memory. A memory chip consists of much less physical material than a book, and therefore it has a much smaller Boltzmann entropy. Considering the way in which computer storage devices have reduced in size in recent years, while at the same time holding ever-increasing amounts of information, it is not too fanciful to suppose that one day soon we will be able to store one bit of information per atom. When this starts to happen, the degrees of freedom used in the two types of entropy will begin to converge.

And if we use the *same* degrees of freedom when we calculate the two types of entropy, we will find that the answers agree precisely. The two entropies will be equal.

This observation has led some physicists to conclude that, at the deepest level, nature is made of information. From this viewpoint, space and time, matter and energy — the concepts we usually think of as being the fundamental properties of the physical world — are secondary to the flow of information. Although some physicists might regard all this as a tenuous conclusion based ultimately upon a confusion of terminology, it is at least interesting to regard the Shannon entropy and the thermodynamic entropy as being equivalent — for the same degrees of freedom — and to see where that leads us. From now on, then, we can talk just of entropy. This leads to a fascinating question, which we can phrase in slightly different ways. What are the *ultimate* degrees of freedom for these entropy calculations? What is the true thermodynamic entropy of a given piece of matter? What is the maximum amount of information that, in principle, a given amount of matter can contain?

You might think that these are difficult questions to answer if we do not know the fundamental structure of matter. Books contain atoms; atoms consist of nuclei and electrons; nuclei contain protons and neutrons; protons and neutrons consist of quarks; quarks are perhaps merely excitation modes of strings. Until we know for certain what the truly fundamental objects of the Universe are, it would seem to be impossible to calculate the true thermodynamic entropy of an object. Remarkably, although such questions may seem to be beyond us, a study of black holes can give us answers.

Practical Thermodynamics

We have to keep reminding ourselves that these theoretical uncertainties do not pose a practical problem for the thermodynamicist. The thermodynamics of steam engines can be perfectly understood without a knowledge of protons or quarks or superstrings: this is because the low-energy conditions inside a steam engine cannot cause any change in the behavior of protons, much less quarks or superstrings. We know which are the appropriate degrees of freedom to use when we discuss the entropy of steam engines. But even though the calculation of true thermodynamic entropy is not a practical problem, it *is* a theoretical problem to which we should have an answer.

As we saw in the previous chapter, Jacob Bekenstein postulated that the entropy of a black hole is proportional to the surface area of its event horizon. Stephen Hawking then showed that the entropy of a black hole is equal to one quarter of its horizon area (when measured in units of Planck area). According to Bekenstein, when an object falls into a black hole the entropy associated with that object does not simply vanish; the mass of the black hole increases and so does the area of the horizon and thus the black hole entropy. Bekenstein conjectured that in *any* process involving a black hole, the sum of the black hole entropy and any entropy outside the event horizon can never decrease. This general form of the second law of thermodynamics always holds. In 1986, Rafael Sorkin of Syracuse University showed that, on very basic grounds, the entropy we calculate in black hole processes must involve the most fundamental degrees of freedom — *whatever those fundamental degrees of freedom happen to be.* So even if we do not know explicitly what is the ultimate structure of matter, we can still calculate the true entropy of a black hole.

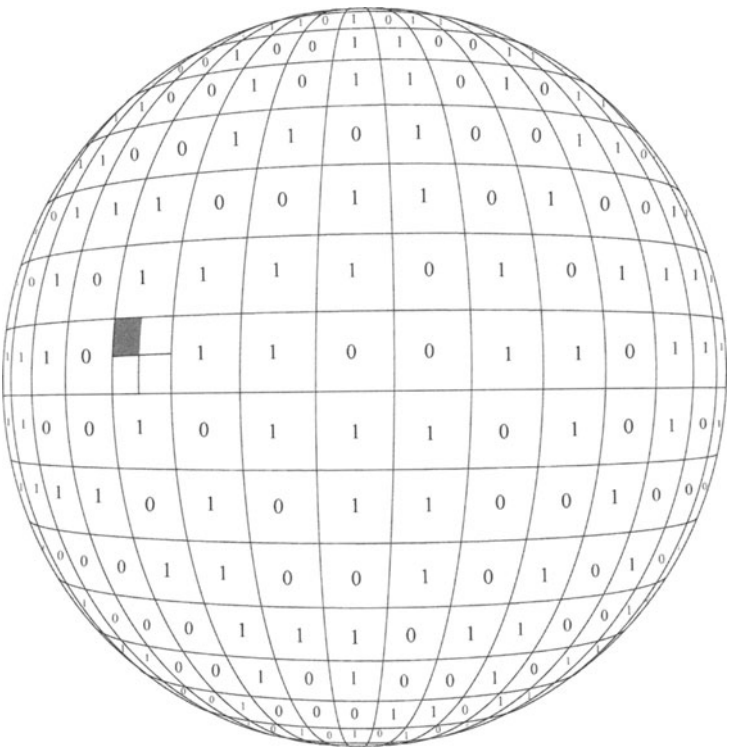


FIGURE 114
We can think of the entropy of a black hole as being information written on its event horizon. Each bit of information corresponds to four Planck areas; the shaded area shown on the sphere represents one Planck area.

All this leads to an astonishing conclusion. Let's consider the following gedanken experiment. Imagine for a moment that the Universe consists

of a three-dimensional lattice. The lattice spacing is the Planck length of 10^{-35} meters, and there is a particle at each lattice point. Each particle can be in one of two states; perhaps the particles are spin-½ fermions, with a spin that points either up or down. (This model is actually not too outlandish; after all, we do not know what happens down at the Planck length. In any case, the exact details do not matter; the conclusions are robust.) Now consider a spherical volume of such a space. We know exactly how many particles are in this volume, and we can count the total number of different states in the volume. In other words, we can calculate the entropy of this volume. It is difficult to avoid the conclusion that the maximum entropy is proportional to the size of the volume. This is just common sense: if you double the volume, you double the number of states that contribute to the entropy. But this answer must be wrong, at least when we talk about a quantum theory of gravity, and it is easy to see why. Suppose that the mass inside the volume is close to gravitational collapse, but is not yet quite a black hole. (There is no event horizon.) The mass fits inside a closed surface of area A; this is not the area of an event horizon, simply the surface area of the volume. Now throw a little extra mass into the system and trigger a gravitational collapse. A black hole will form, and the area of the event horizon will be smaller than A. In turn, this means that the black hole entropy will be smaller than $A/4$, in our usual units. But the total entropy can never decrease; the second law of thermodynamics forbids it. We are thus led to the conclusion that the original entropy of the mass inside the volume could have been no larger than $A/4$. In fact, the entropy of *any* isolated system with a surface area of A must be no larger than $A/4$. This is a completely general result. It does not matter what a physical system is made of, nor what the ultimate constituents of nature turn out to be; the second law of thermodynamics — if it holds for black hole processes — tells us that the fundamental entropy of a system depends upon its surface area, not its volume. Put another way, the maximum amount of information you can pack into a system depends upon the surface area of the system, not the volume. (Remember that we are talking about theoretical limits here. Even with the amazing improvements in computing power of recent decades, it is unlikely that we will reach such a limit on storage capacity any time soon.)

This is a very strange result. Surely there is some way to get around it? Can you not just keep adding information to a system, so that it contains more entropy than is contained in its surface area? Suppose, for example, that you were to make a pile of books. Start with the contents of the British Library, then add the Bodleian in Oxford, and then add the Library of Congress. This would constitute a large amount of Shannon entropy and a vast amount of Boltzmann entropy. Moreover, as you started to make

this pile, the entropy would clearly grow with the volume of the books: double the radius of a spherical pile of books and you increase the volume by a factor of eight, and eight times as many books means eight times as much entropy. The surface area of the pile increases only as the square of the radius, so if you double the radius then the area only increases by a factor of four. So clearly, you will eventually reach a situation where the entropy inside the volume of books exceeds the entropy at the surface area, right? Well, no. Calculations show that just as this situation is reached, the pile of books would collapse under its own weight and form a black hole. The entropy of the hole would be $A/4$ and, if you kept on feeding books into the hole, the area of its horizon would increase in such a way that the general form of the second law of thermodynamics would hold.

There seems no way around it. The maximum amount of information that a physical system can contain depends upon its area, not its volume.

The Holographic Principle

To see a world in a grain of sand
William Blake, *Auguries of Innocence*

It was a consideration of information and black hole behavior, combined with the difficulty of formulating a quantum theory of gravity, that led 't Hooft to the holographic principle. Before stating the principle, it may be worth reminding ourselves what a hologram is (see the boxed text on page 233).

This finding — that the maximum information storage capacity of a black hole depends upon its area, not its volume — so impressed 't Hooft that he made a radical proposal. He said that any plausible quantum theory of gravity must be able to describe all the phenomena happening *within* a volume by degrees of freedom that live on the *surface* of that volume. He called this the *holographic principle*, because it was reminiscent of the way in which all the information required to describe a three-dimensional image is encoded in a two-dimensional hologram.

The holographic principle is a truly bizarre and counterintuitive idea. It says that all the physics occurring *in* a particular three-dimensional region (your living room, for example) can be described by a theory that resides only *on* the two-dimensional boundary of the region (the flat walls, floor, and ceiling of your living room). Of more relevance to a theory of quantum gravity, it says that all the information about the contents of a black hole is stored on a two-dimensional surface just outside the event horizon, as shown in Figure 116.

Holograms

Holography was developed in 1947 by Dennis Gabor, who was then working at the British Thomson–Houston Company Research Laboratory in Rugby, England. Gabor won the 1971 Nobel prize for physics for his invention.

Nowadays, holograms are everywhere. Banknotes carry them to deter forgers; advertisers use them as a cheap way of attracting attention to a product; art shops sell striking holographic images for no more than the cost of a two-dimensional print. But do not let the ubiquity of holograms blind you to their remarkable properties. A hologram captures all the information about how an object looks: a hologram is a photograph that can capture a three-dimensional image on a two-dimensional surface. (It was Gabor who coined the word *hologram*. It comes from the Greek words *holo*, meaning "whole," and *gramma*, meaning "message.")

A photograph records the intensity of the light reflected by an object. A point on the object that is white reflects more light than a point on the object that is black. Thus the "white" point exposes more silver halide in a photographic film than the "black" point. The combination of millions of such points, each of which records a particular intensity of light, is ultimately recognized by us as a two-dimensional representation of the object.

A hologram works in quite a different way. It records the phase information of light that has been reflected by an object. Holography works by illuminating an object by a beam of light called a *scene beam* (also known as an object beam). The physical characteristics of the object change the intensity and phase of the scene beam. The reflected scene beam is then allowed to interfere with a *reference beam*, which has the same characteristics — the same wavelength and intensity — as the original scene beam. The interference pattern is then recorded on film to form the hologram.

All the points on the object that reflected light which interfered with the reference beam are focused to their respective positions in three-dimensional space. The front of the object appears to be closer, the back appears to be farther away, and the midpoints appear in between. Thus a three-dimensional image appears on the two-dimensional surface of the film.

For this method to work well, we need highly coherent light, of the sort produced by lasers. It is a mark of Gabor's genius that he produced the first hologram with non-coherent light, more than a decade before the laser was developed.

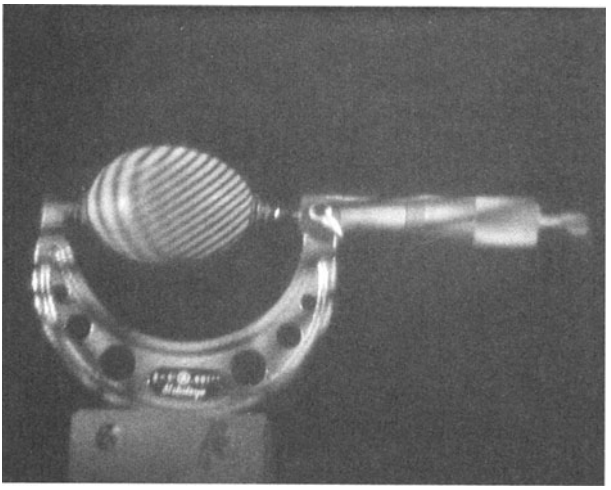


FIGURE 115 A double-exposure hologram of a micrometer clamping an egg. The image is striking (when seen in real life), but holograms also have a serious scientific purpose. The fringe patterns on the egg occur when it is placed under pressure from the micrometer; movements as small a wavelength of light show up as line or fringe. Such techniques are useful for showing how objects respond to stress.

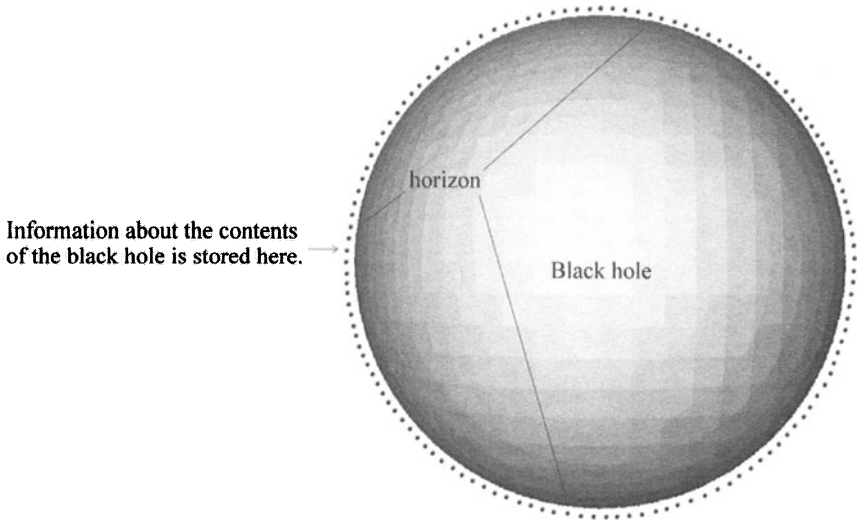


FIGURE 116 An astrophysical black hole — such as the kind found in the center of galaxies or after the gravitational collapse of a massive star — is a three-dimensional object. But if the holographic principle holds true, all the physics that occurs inside the hole's event horizon can be described by a theory that is defined purely on a two-dimensional surface just outside the horizon. Information in two-dimensions is sufficient to describe three-dimensional physics, rather like a hologram.

The holographic principle says that the world is essentially two-dimensional, but the two-dimensional surface contains information that somehow is encoded in such a way that we recover our usual three-dimensional world. Or there is another possibility: perhaps *we* are the holograms, living on the boundary surface of a five-dimensional spacetime;

FIGURE 117 Raphael Bousso was a student of Stephen Hawking at Cambridge University. He is now based at the University of California, Berkeley.

if the holographic principle is correct, we would not notice. We would be like the prisoners in Plato's cave, convinced that the world was "flat" and oblivious to the extra dimension.

Soon after 't Hooft and Susskind made their initial investigations of the holographic principle, Raphael Bousso, a student of Stephen Hawking who is now based at the University of California, Berkeley, modified the principle in a way that resolved some problems in the original formulation. But it was Juan Maldacena's celebrated paper of 1997 that really made physicists sit up and take notice.

THE ADS/CFT CORRESPONDENCE

You start with brane and the brane is BPS.
Then you go near the brane and the space is AdS.
Who knows what it means? I don't, I confess.
Ehhhh! Maldacena!
Jeff Harvey, "*The Maldacena*"
(To be sung and danced to the tune of "The Macarena")

If a realistic quantum theory of gravity must exhibit the holographic principle, as 't Hooft suggested, and if M-theory is indeed a theory of quantum gravity, it follows that M-theory should exhibit the holographic principle. It does.

Juan Maldacena is a young Argentinean physicist who is now based at Princeton's Institute for Advanced Study (also home to Ed Witten and Nathan Seiberg). As Maldacena pondered the success of the D-brane description of black holes, he realized that there were two ways of looking at the calculations. D-branes are massive objects, and so they curve the spacetime in which they are embedded. From this point of view, gravity

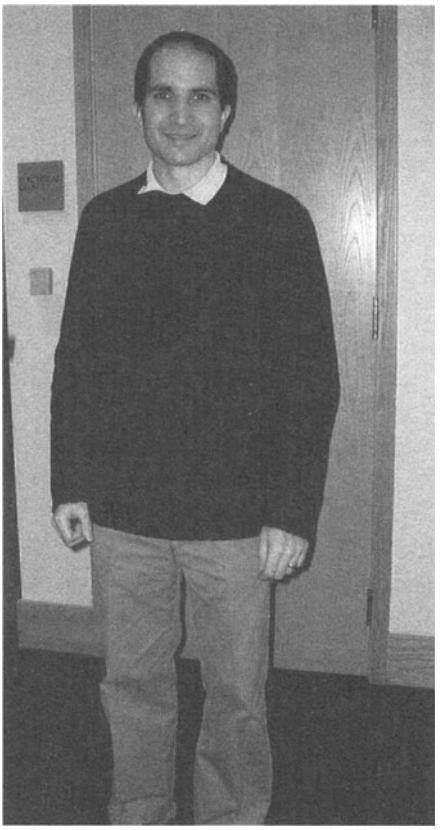


FIGURE 118 Juan Maldacena was born in Argentina; he is now a professor of physics at the Institute for Advanced Studies in Princeton. His 1997 paper on the AdS/CFT correspondence is one of the most highly-cited papers in the recent history of particle physics.

is important. On the other hand, the quantum theory of D-branes is a gauge theory. These two viewpoints suggested to Maldacena that there was a link between gauge theories and string theories. The tangle of branes and strings might be describing not only the quantum states *inside* the hole but also the geometry of spacetime *around* the hole near the event horizon. This looked suspiciously like a holographic idea, since it would be linking physics inside a volume with physics defined on the area of that volume. These musings led Maldacena to conjecture the existence of a precise relationship between a theory with gravity acting in a certain spacetime and a quantum field theory without gravity acting on the boundary of that spacetime. It was a concrete example of the holographic principle, and it came from M-theory.

Maldacena's conjecture goes by the name of the AdS/CFT correspondence. To understand what the conjecture says, we need to make a brief digression.

Let's start with the AdS part of the correspondence.

Anti-de Sitter (AdS) spacetime

The "AdS" stands for *anti-de Sitter spacetime*, a particular type of spacetime that has long been studied by theorists. It is named after Willem de Sitter, a Dutch astronomer who, in 1917, found a particular solution to Einstein's equations of general relativity. The solution described a highly symmetrical but rather peculiar type of spacetime — the so-called *de Sitter spacetime* — in which the universe was completely devoid of matter but which contained a positive cosmological constant. (See the boxed text on page 239 for a brief discussion of the importance of the cosmological constant.) De Sitter found that this empty spacetime expands at an ever increasing rate. You might well wonder in what sense a universe can be said to "expand" if it contains no matter. What expansion means in this context is that if we introduce two test particles — objects with negligible mass, so that there is negligible gravitational attraction between them — the distance between the particles increases with time. In fact, in the de Sitter spacetime, the distance increases at an exponential rate. The de Sitter spacetime is a rather simple theoretical model, and there seems to be no obvious reason why such a strange construct — a highly symmetrical, empty, expanding spacetime — should be relevant a world such as ours (which contains matter that is distributed, at least on small distance scales, in a 'lumpy' and asymmetrical fashion). But one of the most intriguing discoveries of recent science, made 80 years after de Sitter published his model, is that the expansion of our Universe appears to be accelerating. In the far future our Universe might be approximated by a de Sitter spacetime.

The cosmological constant in de Sitter spacetime is positive. If we make the constant *negative*, so that the repulsion becomes an attraction, we have an anti-de Sitter spacetime: AdS for short.

As with the original de Sitter space, AdS is a simple and symmetrical solution to Einstein's equations. (It is, though, even less realistic than de Sitter space as a model of our Universe since not only is it devoid of matter, the cosmological constant is negative.) It is difficult to visualize anti-de Sitter space, but in some ways, we can think of it as the opposite of a sphere. A sphere is a positively curved surface. By this we mean that the interior angles of any triangle drawn on the surface of a sphere add up to more than 180°, and parallel lines drawn on the sphere always meet. (Compare this with a Euclidean plane, which has zero curvature. The interior angles of a triangle drawn on a plane always add up to precisely 180°, and parallel lines never meet.) Well, AdS is a *negatively* curved surface: the interior angles of a triangle drawn on the surface always sum to less than 180°, and there are an infinite number of lines parallel to a given line that go through a fixed point. A saddle, for example, has a negative curvature. The sphere

has the symmetry property that no point on the surface is different to any other point — it has a constant positive curvature. AdS has the symmetry property that its negative curvature is the same everywhere.

An AdS spacetime possesses a strange type of boundary: it is located at infinity. The equations of general relativity tell us that, because of the negative curvature, a massive particle cannot reach the boundary. But photons can. If you send out a photon, the photon can reach infinity in a finite time; if it is reflected at the boundary, it will get back to you in a finite time. The length of time it takes to reach the boundary depends upon the value of the cosmological constant. If we lived in an AdS universe with a sufficiently negative cosmological constant, a flash of your car headlights could reach the edge of the universe and bounce back to dazzle you before you had a chance to change gears. In this sense, then, AdS possesses a natural and well-defined boundary. This boundary property of AdS is interesting from the holographic point of view, since holography also requires a boundary, or screen.

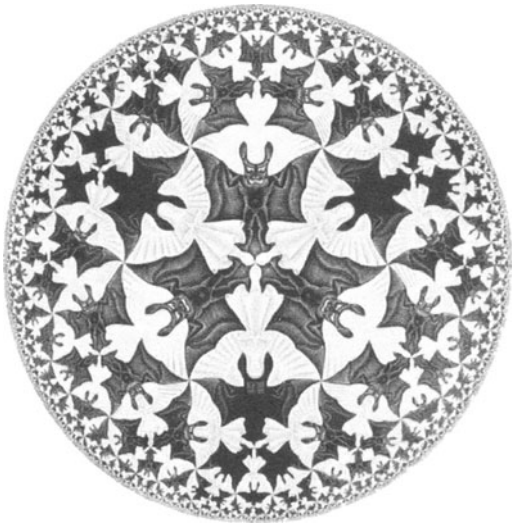


FIGURE 119 Escher's famous woodcut Circle Limit IV, also known as Heaven and Hell, is a representation of a negatively curved geometry. Each of the angels and devils that live here in the "bulk" is actually the same size; distances stretch as one approaches the circumference, and the negative curvature of the space causes distortion. A group of such images, stacked together to form a tower and with time increasing as one moved up the tower, would represent a snapshot of an anti-de Sitter spacetime.

Maldacena considered type IIB string theory in ten dimensions and then compactified five of the dimensions on a five-dimensional sphere. The remaining five-dimensional spacetime was anti-de Sitter, AdS_5. In other words, at each point in the five-dimensional anti-de Sitter spacetime there was a five-dimensional sphere. The superstrings lived in the full ten-dimensional theory, but the interesting physics occurred in AdS_5. So in AdS_5 we have a string theory of gravity living in a five-dimensional spacetime with a well-defined boundary.

The Cosmological Constant

As Einstein developed general relativity, he realized that the equations of his theory — which relate spacetime and matter — should describe how the Universe evolves over time. The solutions of his equations implied that the Universe must collapse, since gravity is attractive. But in 1917, when Einstein was working on his theory, people believed the Universe was static. To solve this problem, Einstein introduced a "cosmological constant" into the equations of general relativity. The term was entirely consistent with all the principles of his theory, so he was quite at liberty to do this. Indeed, we now know that there is no good reason *not* to include such a term in the equations from the outset.

The cosmological constant could be positive or negative. A positive constant would produce a long-range repulsive force that countered the tendency of the Universe to collapse. It would describe a type of antigravity — the energy of which depends upon the presence of empty space.

In the 1920s, astronomers began to realize that the Universe was not static; it is expanding. There was no need for a repulsive force, since the universal attraction of gravity could be acting to stop, or at least slow, the expansion. Einstein dropped the cosmological constant from his equations and called its introduction his "biggest blunder."

We now know that quantum mechanics can endow the vacuum of empty space with energy, and this energy can act in such a way that it mimics exactly the cosmological constant that Einstein introduced. (Quantum mechanics predicts the existence of "zero point energy." As we know, particle–antiparticle pairs are constantly being created out of the vacuum, and although they exist for only a short time before annihilating they endow the vacuum with a non-zero potential energy.) So a cosmological term occurs naturally, and it thus becomes a matter for theorists to predict its value or explain the experimentally observed value.

From observations, astronomers know that the cosmological term is small and positive, or possibly zero. (It must be small, because if it were large and positive the Universe would have blown itself apart long ago under the repulsive force; if it were large and negative the Universe would have long ago collapsed under the attractive force.) One of the abiding problems of theoretical physics is that when we calculate the value of the cosmological constant using our best theories, we end up with an answer that is about 10^{120} orders of magnitude larger than the experimentally observed bound! This has been called the worst mismatch between theory and experiment in all science. To date, no one even has a good theoretical framework for understanding how to reconcile this discrepancy.

It turns out that a theory of gravity in AdS_5 is comparatively easy to work with. In fact, because the symmetry and simplicity of AdS makes for easier calculations, Maldacena was able to deduce many results for the theory in this five-dimensional volume.

Conformal field theory (CFT)

The boundary of five-dimensional AdS spacetime is a four-dimensional surface (just as the boundary of a three-dimensional volume is a two-dimensional surface). Inspired by the holographic principle, Maldacena suggested that this string theory of gravity in five dimensions was dual to a quantum field theory of point particles living on the four-dimensional spacetime boundary. The particular field theory that he could identify as being dual was an SU(N) gauge theory with $\mathcal{N} = 4$ supersymmetry, in the limit of large N. This four-dimensional field theory possessed another symmetry: it had no preferred energy scale, so the theory looked the same at all energies. It was a conformal field theory (CFT).

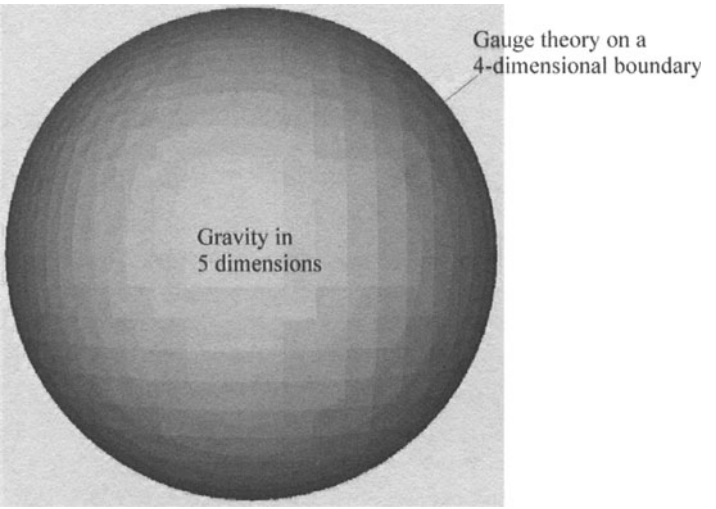


FIGURE 120 The AdS/CFT correspondence says that a particular theory containing gravity in a five-dimensional spacetime can be described by a gauge theory without gravity that is defined on the four-dimensional boundary of the spacetime.

So the Maldacena conjecture was of an AdS/CFT correspondence. A superstring theory of gravity in a five-dimensional anti-de Sitter spacetime was holographically dual to ordinary quantum field theory without gravity acting on the boundary of the spacetime. The quantum field theory was

in some sense a "hologram" of the full gravitational theory. Gravity was in some sense the same as field theory. This was the first concrete realization of the holographic principle. And it was an astounding conjecture.

Soon after Maldacena announced his conjecture, three Princeton University theorists — Steven Gubser, Igor Klebanov, and Alexander Polyakov — published a paper that provided clear recipes about exactly what one could calculate using the AdS/CFT correspondence. Just four days after the publication of this paper, Witten published an independent analysis that came to the same conclusion. The conjecture was on firm ground.

There is a problem if we try to learn something about our real world using Maldacena's original formulation of the AdS/CFT correspondence. The five-dimensional AdS space is empty and highly symmetric, which is not at all like the messy radiation-filled four-dimensional Universe we observe. And the CFT is a supersymmetric gauge theory that has no preferred energy scale, which again is not like our real world. The real world, at least at the low energies we work with, seems not to possess supersymmetry, and the real world certainly has preferred energy scales and is thus described by *non*-conformal gauge theories such as QCD. Nevertheless, the Maldacena correspondence is so remarkable that physicists have worked hard to extend it to different types of spacetime and gauge theory. Extending the correspondence so that it includes realistic theories is an active area of research.

It is not difficult to appreciate the enthusiasm with which theorists greeted Maldacena's conjecture. Not only was it saying that gravity was in some deep way the same as quantum field theory and also implementing the holographic principle in a concrete way, it also shed light on other aspects of physics — large-N theories, for example.

Tying Strings Together: AdS/CFT and Large-N

We shall not cease from exploration
And the end of all our exploring
Will be to arrive at where we started
And know the place for the first time.
T. S. Eliot, *Four Quartets, 'Little Gidding'*

In what must be one of the most tortuous routes in science, Maldacena's conjecture meant that string theory had come full circle.

Let's recap where the string idea came from. Historically, string theory arose out of attempts to find a theory of the strong interaction, the inter-

action that occurs between quarks in hadrons. A string model arises quite naturally and gives an intuitive feel to several key features of the strong interaction. For example, we can think of a meson as a quark and an antiquark at the opposite ends of a string. If we try to isolate a quark, perhaps by using a collider to hit it with a quark from another hadron, the string becomes stretched. As the string stretches, the energy contained in the string increases. Before the quarks are separated by macroscopic distances, the string will break. If we were hoping to see an isolated quark and an isolated antiquark, quantum mechanics foils us: there is always enough energy in the stretched string to create another quark–antiquark pair. The newly created quark pairs up with the existing antiquark, and the newly created antiquark pairs up with the existing quark. We now have *two* mesons, each consisting of a quark–antiquark pair at the end of a string. The string model thus gives us a rough-and-ready picture of quark confinement. It also provides a picture of various other aspects of the strong interaction. So hadronic physics gave a natural birth to string theory.

By 1973, however, physicists had a quantum field theory of the strong interaction: quantum chromodynamics. It was a gauge theory based on SU(3), with the "3" simply meaning that the quarks come in three colors. The sometimes nebulous arguments of the string model gave way to a precise and well-understood conceptual framework. Unfortunately, although physicists now had a theory of the strong interaction, there were many situations in which they could not solve the theory; it was impossible to prove quark confinement, for example. No one doubted that QCD was the correct theory with which to analyze confinement in principle, but in practice the mathematics were too difficult for anyone to address such questions analytically. The usual perturbative approach failed.

String theory may come to the rescue of quantum chromodynamics in these difficult-to-handle, non-perturbative regimes. There are two strands to this story, and the AdS/CFT correspondence discovered by Maldacena ties them both together.

The first strand of the story starts in 1974, when 't Hooft came up with perhaps the best suggestion anyone has yet made about how to calculate the most interesting aspects of strong interaction physics (quark confinement, the hadron masses, and so on). These interesting aspects also happen to be the most intractable items for QCD. The idea 't Hooft had was that QCD would be much simpler if quarks came not in three colors, but in N colors, where N is a very large number. So he proposed studying a gauge theory based on SU(N) rather than SU(3) as a simple approximation of QCD. As 't Hooft developed this idea, he noticed that the large-N gauge theory behaved in many ways like a string theory. So he made the conjecture that SU(N) gauge theory in four dimensions — the simpler version of QCD —

was equivalent to a string theory. If the conjecture were correct, it would explain why a string model had been successful in describing many aspects of hadronic physics. Furthermore, 't Hooft showed that the string coupling constant would be proportional to $1/N$, so at large N we could employ the perturbation approach to studying string interactions. The hope was that the large-N approximation could be solved, and one could work back to the real-world case of $N = 3$. It was a good idea, but unfortunately the simplification that came from assuming a large number of colors was not enough for 't Hooft to perform realistic calculations. Progress on the large-N approximation slowed down — but the idea that QCDmight be equivalent to a string theory never entirely died.

The second strand of the story is the one outlined in this book. Physicists discovered that string theory had the capacity to encompass *all* the fundamental interactions, including gravity. For this to work, however, the theory had to incorporate supersymmetry; moreover, a perturbative approach to the theory could be defined consistently only if the number of spacetime dimensions was ten. In fact, five different consistent superstring theories could be defined in a ten-dimensional spacetime. Further work showed that a consistent theory of strings required the presence of branes — subspaces of a higher-dimensional space — as well as strings. The study of branes led to the discovery of duality relations between the various superstring theories; the five superstring theories were different manifestations of a unique unified theory — M-theory — that could be defined in 11 dimensions. The discovery of D-branes was of major importance; within the M-theory framework, a study of D-branes led to the first major advance in understanding the microscopic properties of black holes. Insights gained from the theoretical analyses of black hole properties, combined with the holographic principle advocated by 't Hooft and Susskind, led Maldacena to conjecture a correspondence between a string theory of gravity and a gauge theory. And the gauge theory was a supersymmetric theory with gauge group SU(N), where N was large. It was a large-N theory!

More than two decades after the 't Hooft's suggestion, Maldacena had identified a connection between a large-N gauge theory and a string theory. The route to Maldacena's conjecture was byzantine — but the fact that a route exists at all seems to be yet another indication of some deep underlying structure at work.

For the first time, we may be on the verge of moving backwards *from* a unified theory (M-theory) *to* the Standard Model of particle physics. Rather than groping blindly towards an unknown theory, physicists can use the duality between gravity and gauge theory to calculate quantities in QCD that are otherwise impossible. A calculation that is intractable in QCD can

be straightforward in the corresponding picture of gravity. For example, calculations of the masses of glueballs, which are difficult to do analytically in QCD, are much easier to do in the dual gravity picture. In recent years, physicists have started out on this approach, and the results are encouraging. Most of this work is in gauge theories that are close, but not quite identical, to real-world QCD. But even this caveat is weakening.

String theory started out as a theory of quark interactions and in that it failed. But by becoming a theory of gravity, string theory is at last becoming useful as a theory of quark interactions!

11

WHEN WORLDS COLLIDE

I think all you scientists are crackpots.
Sydney Stanton, *When Worlds Collide*

Perhaps the key untested idea in all of physics is that extra spatial dimensions might exist. Once we admit the possibility of extra dimensions several avenues for exploring theories of unification open up. Without extra dimensions there simply does not seem to be room for unification of the various forces. But where are these extra dimensions? For six decades after the pioneering work of Klein, the belief was that extra spatial dimensions — if they exist — are unobservable because they are tiny. The assumption was that any extra dimensions must be curled up on the Planck length scale of 10^{-35} meters. Eventually, that assumption began to be challenged. In 1983, for example, two physicists at Moscow's Institute for Nuclear Research, Valery Rubakov and Misha Shaposhnikov, suggested there could be large extra dimensions that we have not yet observed. And in 1990, Ignatios Antoniadis, who was based at CERN, gave a serious

proposal for how some of the extra dimensions of string theory could be 10^{-19} meters in length — still small, for sure, but much larger than previously thought. Such proposals were not without drawbacks, however, and in any case the ideas were not motivated by any deep underlying principles. Then came M-theory.

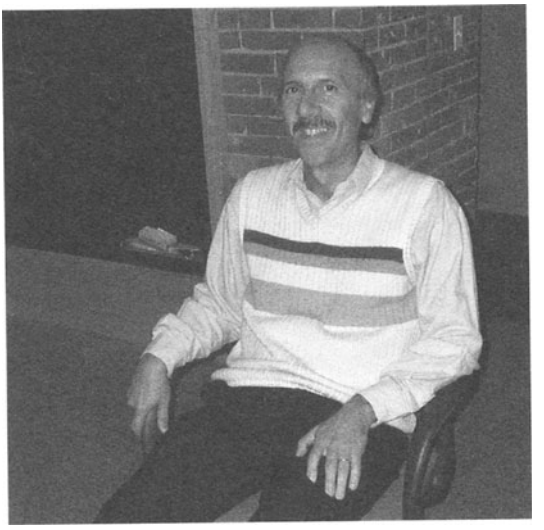


FIGURE 121 Ignatios Antoniadis is a member of the Theory Division at CERN. He was one of the first to suggest that the extra dimensions of string theory might be much larger than the usual Planck scale.

The mysterious and still undefined M-theory, whatever it may turn out to be, is a theory that makes sense only if nature contains multi-dimensional branes as well as one-dimensional strings. Once branes are allowed, there are reasons for believing that extra dimensions could be macroscopic. These reasons *may* turn out to be weak, and M-theory does not require the existence of large extra dimensions. But we are certainly free to postulate their existence.

Theorists have used the ideas of branes and large extra dimensions to take a fresh look at many old problems. In cosmology, for example, the ideas allow physicists to cast an ancient belief into a modern form: perhaps our Universe had no beginning and will have no end. Some physicists have theorized that our world periodically collides with another world — two branes (each world exists on one) clashing throughout eternity, with each collision giving rise to an event we call the Big Bang. Ideas like these are at the speculative reaches of physics — many of them will no doubt turn out to be wrong — but they are powerful ideas, nonetheless. Scenarios involving branes give rise to the most exciting notion in modern physics: that some day soon, perhaps even by the time you read this book, we will be probing the higher dimensions in experiments.

BRANE NEW WORLD

O brave new world.
William Shakespeare, *The Tempest*

When superstring theory became popular, after the First Superstring Revolution of 1984, the common assumption was that the six extra spatial dimensions required by the theory were curled up on scales comparable to the Planck length. Since the Planck length was 10^{-35} meters, this was bad news for physicists who wanted to try to detect the extra dimensions. The extra dimensions would be far too tiny to be observable at the low energies of our everyday world. Nevertheless, the assumption that the extra dimensions were Planck-sized was just that: an assumption. It is certainly worth asking about the *experimental* limits on the size of the extra dimensions.

We saw on page 148 that a key prediction of Kaluza–Klein theory is the existence of a "tower" of particles. According to quantum mechanics, particles possess wavelike aspects and these waves can wrap themselves around the extra dimensions. Such behavior gives rise to an infinite number of higher-energy versions of the original particle, so-called Kaluza–Klein particles. (The same behavior can happen in string theory.) Now, the separation in energy of the Kaluza–Klein particles depends upon the size of the extra dimensions. So if there were an extra dimension with a radius of 10^{-19} meters, for example, then the energy separation between particles in the tower would be 1 TeV. The Z^0 particle, which has a mass of 91 GeV, would have Kaluza–Klein cousins of mass 1091 GeV, 2091 GeV, 3091 GeV, and so on. The fact that no Kaluza–Klein particles have been observed in collider experiments puts a bound on the size of the extra dimensions: they can be no larger than 10^{-18} meters.

The experimental bound of 10^{-18} meters may be much larger than the Planck scale, but it is still suggests that any extra dimensions are likely to be too small for us to probe directly. At least that is what people believed until 1995. The experimental limits on the size of the extra dimensions depends upon the tools we use to probe them. Polchinski's discovery of D-branes gave physicists fresh insight into this problem and thus began to change how people viewed the extra dimensions.

D-branes, as we saw in an earlier chapter, are multi-dimensional membranes that are free to move in a higher-dimensional space. (Think of one of those thin, flexible, diamond-shaped kites: it is thin enough to be considered as a two-dimensional object, and it is free to move in the three dimensions of the atmosphere. The kite is in some ways like a D2-brane embedded in a three-dimensional space.) The study of D-brane properties opened up whole new areas of research, but the important property of D-branes in this context is that they are surfaces upon which open strings

are trapped. This is important because the particles of the Standard Model — the quarks, leptons, and gauge bosons — are believed to be excitations of open strings. So if we happened to live on a brane, there would be well understood reasons why the Standard Model particles should be confined to the brane. Gravitons, on the other hand, are *closed* strings; they are free to move away from the brane. (Think of the kite floating in a gentle shower of rain. Most water droplets on the kite can move only on the kite's surface — they are confined to a two-dimensional brane. Such droplets would correspond to the particles of the Standard Model. Some droplets may fall off the kite and are free to move through the atmosphere; these droplets would correspond to gravitons.)

What does this have to do with the possibility of large extra dimensions? Well, part of the work that initiated the Second Superstring Revolution was the investigation by Hořava and Witten of the strong coupling limit of the E(8) × E(8) heterotic string: the superstring theory that was once touted as being the most promising from the viewpoint of unification. As we saw earlier, when the E(8) × E(8) heterotic string is in the strong coupling regime, it develops a new dimension and becomes an 11-dimensional theory. But the new dimension is not a circle. The new dimension is bounded by two walls. (See Figure 122.)

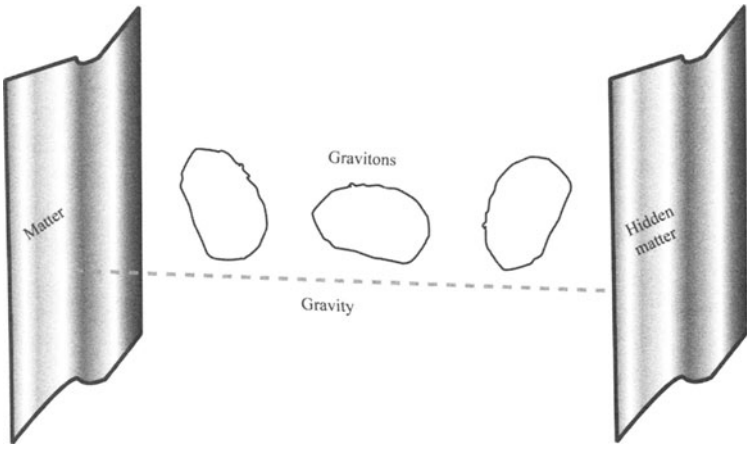


FIGURE 122 A representation of an 11-dimensional spacetime first envisaged by Hořava and Witten. The two planes have three large spatial dimensions — they represent our usual four-dimensional (3 + 1) spacetime. The six extra dimensions of superstring theory are compactified on a Calabi–Yau manifold and are unobservable. Gravitons, which are closed strings, are free to move in the (4 + 1)-dimensional bulk between walls. Thus gravity is aware of the bulk; the fields of the Standard Model are confined to the brane.

In the Hořava–Witten picture, all the Standard Model fields would be restricted to move on a wall, a spacetime with three spatial dimensions; the

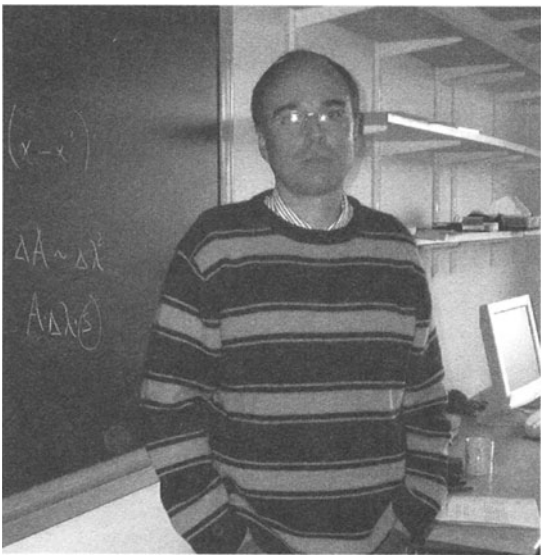


FIGURE 123 Petr Hořava, a native of the Czech Republic, is presently at the Center for Theoretical Physics at the University of California, Berkeley. His research into the physics of branes has been extremely influential.

other six spatial dimensions of string theory would be compactified on a Calabi–Yau manifold to unobservably small dimensions. The other wall would have a similar structure, with gauge fields confined to its surface. Gravity would be free to move in the extra eleventh dimension, in the *bulk* space between the two walls. Hořava and Witten showed that if this eleventh dimension was relatively large — about 10^{-30} meters — one could achieve supersymmetric unification of the three forces with gravity at a length scale of 10^{-32} meters. (It may seem perverse to say 10^{-30} meters is "relatively large," but remember that it is ten thousand times larger than the Planck distance, which is the scale of the compactified dimensions.)

The Hořava–Witten spacetime is a bizarre one; if it were an accurate model of our Universe, it would mean that there is a parallel Universe to our own, a Universe that is just 10^{-30} meters away. Yet despite its proximity we would never see our fraternal Universe, because photons — and all other particles we could use as probes — are trapped within our Universe. Only gravity escapes. (See Figure 124.)

The extra dimension of the Hořava–Witten theory may be bigger than the Planck scale, but at 10^{-30} meters it is still unobservably tiny. The experimental bound when they published this work, remember, was 10^{-18} meters. But this approach to extra dimensions can be adapted. In 1998, three researchers — Nina Arkani-Hamed who was then at Stanford Linear Accelerator Center, Savas Dimopoulos of Stanford University, and Georgi Dvali who was then at the International Center for Theoretical Physics, Trieste — followed on from work they did with Ignatios Antoniadis to show how an extra dimension could be huge.

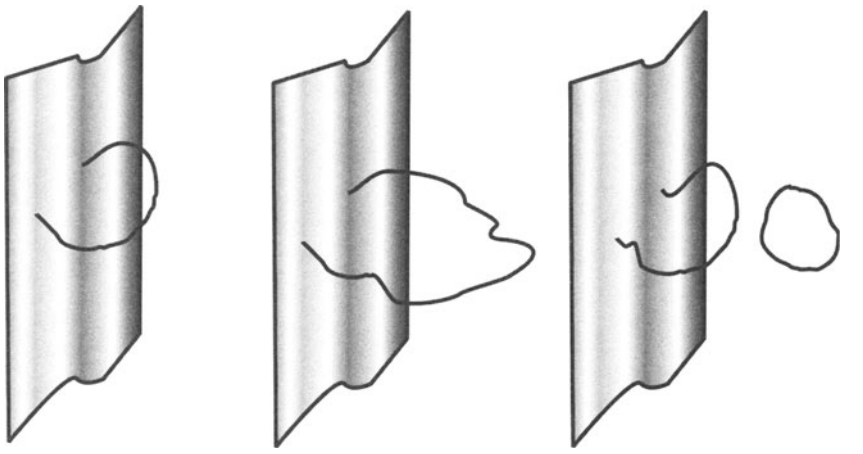


FIGURE 124 Open strings (vibrations of which give rise to the familiar particles like electrons, quarks, and photons) are anchored to the brane. They cannot be pulled off the brane. Closed strings, which correspond to gravitons, are free to move off the brane.

In the discussion that follows it is worth remembering that, although the picture that Arkani-Hamed, Dimopoulos, and Dvali sketched may have been inspired by M-theory, it is independent of M-theory. Their picture might be correct even if M-theory is false. Similarly, M-theory might turn out to be the ultimate theory of nature, yet the extra dimensions that are required by the theory might well be unobservably small.

The idea of Arkani-Hamed–Dimopoulos–Dvali (ADD, for short) starts with the notion that our Universe is a 3-brane — it has three spatial dimensions plus time — embedded in some higher-dimensional space. The Standard Model particles are anchored to the brane and are thus unable to probe the higher dimensions. Gravity, on the other hand, must exist throughout the full higher-dimensional space. Classically, general relativity says this must be the case. According to M-theory, gravitons must propagate throughout the full space, throughout the bulk in other words, because the graviton is a closed string. The key point of the ADD idea is this: *we have only ever attempted to probe extra dimensions by using Standard Model particles.*

When we argued on page 148 that experiments had ruled out extra dimensions that were larger than 10^{-18} meters, we were tacitly assuming that the strong and electroweak gauge interactions could live everywhere in the full 11-dimensional spacetime. But if the Standard Model is confined to the brane, and only gravity is aware of the extra dimensions, we have a *very* different picture (see Figure 126). We have not studied gravity in anything like the detail with which we have studied the other fundamental

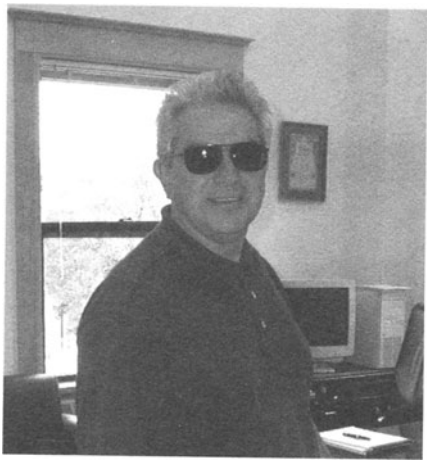


FIGURE 125 Savas Dimopoulos is a professor of physics at Stanford University. He was one of the developers of the Supersymmetric Standard Model — a leading candidate for physics beyond the Standard Model — and was one of the developers of the idea of large extra dimensions.

interactions; gravity is so feeble that it is difficult to test it precisely. So the question becomes: if only gravity is aware of extra dimensions, how large can extra dimensions be given that we have not noticed them?

In their 1998 paper ADD gave a remarkable answer to this question: extra dimensions could be as large as 1 millimeter without violating any of the experimental results then in place! No one had ever even detected the gravitational attraction between millimeter-sized objects, much less investigated the details of the gravitational force on these scales.

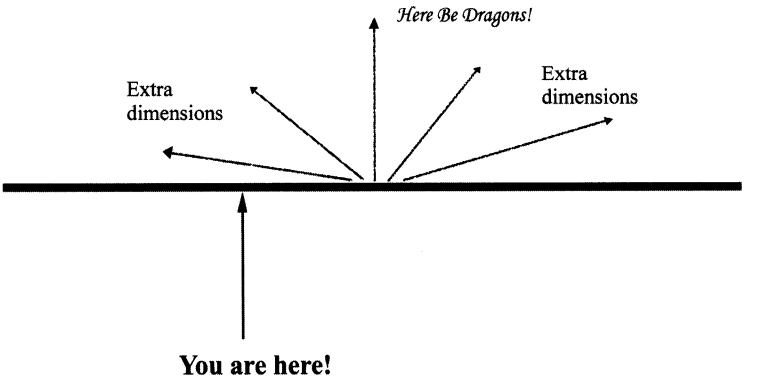


FIGURE 126 In the picture developed by Arkani-Hamed, Dimopoulos, and Dvali, everything that physicists have discovered, and everything that we think we know about, is confined to a thin slice of the Universe. Anything away from this slice is unexplored territory!

It seems impossible to believe that an entire millimeter-sized dimension could have gone unnoticed all this time, but if it is a gravity-only dimens-

ion then we can only investigate it using gravity. And the very weakness of gravity means that it is exceedingly difficult to design experiments to test gravity at short range. As we shall soon see, the experimental community was energized by the thought of detecting extra dimensions. In 1998, however, there was no experimental proof that gravity even existed at distances less than a millimeter, so ADD were certainly free to postulate such large gravity-only dimensions.

If there are indeed extra dimensions that are available only to gravity, the behavior of gravity will be modified on distance scales that are comparable to the size of these dimensions. Gravity will increase rapidly in strength at small distances, and Newton's inverse square law of gravity will no longer hold (see the boxed text on page 253). In their paper, ADD emphasized that "small" distances in this context meant 1 millimeter, since experimenters had not tested the inverse square law at the submillimeter range. These distances were incomparably larger than the usual distance scale of string theory. One could therefore reasonably talk of the possibility of *large* extra dimensions. Extra dimensions — dimensions that are hidden to sight, smell, or touch yet are closer to you than the shirt on your back — could be as large as the nib of a pen.

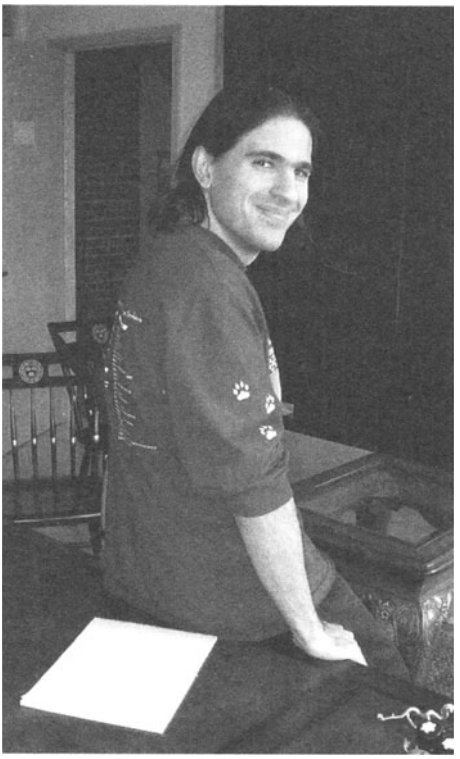


FIGURE 127 Nima Arkani-Hamed is a professor of physics at Harvard University. Much of his research has been driven by the attempt to solve the hierarchy problem.

Modifying Gravity at Small Distances

Suppose that, in addition to the usual three dimensions of space, there is a single extra dimension that it is curled up into a small circle of radius R. Now consider the gravitational field that arises from a point mass.

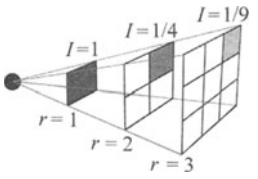


FIGURE 128 The inverse square law. The gravitational force, I, decreases as the square of the distance, r, from a mass.

At distances larger than R, the gravitational field lines from the mass spread out uniformly in the three dimensions of normal space. Since the surface area of a spherical shell centered on the mass increases as the square of the distance from the mass, the gravitational force is diluted at that rate. For large distances, r, the gravitational force goes as $1/r^2$. This is Newton's inverse square law of gravity, and it has been thoroughly tested for large distances. (We know it works well for orbits in the Solar System.) The inverse square law is a direct consequence of living in a Universe with three large spatial dimensions. But what about short distances, when r is less than R?

At distances very close to the mass, the gravitational field lines would be free to spread uniformly throughout the *four* spatial dimensions. So close to the mass the gravitational force would obey an inverse cube law and be proportional to $1/r^3$. Therefore, with increasing separation from the mass, the gravitational force would be diluted more quickly than we are familiar with. Equally it would increase in strength with increasing proximity more quickly than we are used to. At short distances, therefore, gravity would be much stronger than we previously thought. If there were n extra spatial dimensions, all curled up with a radius R, then at distances smaller than R gravity would obey a $1/r^{n+2}$ law. At the ultra-large distances of our everyday world, we would be blissfully unaware of any modifications to Newton's inverse square law; the Moon would continue in its stately Newtonian orbit, irrespective of the existence of extra dimensions. But at small distances, gravity would rapidly increase in strength.

The ADD idea of large extra dimensions was not only interesting and exciting, it held out the promise of enabling experimenters to directly probe the unification of gravity with the other forces. That was something no one had dreamed possible; the received wisdom for over a century had been that gravity was inherently much weaker than the other forces. The gravitational attraction between two electrons, for example, is about 10^{43} times weaker than their electrostatic repulsion. Nowadays, the extreme weakness

of gravity is usually stated in the following terms: there are two fundamental scales in nature — the electroweak scale (about 1 TeV, or 10^{-19} meters) at which electromagnetism becomes unified with the weak interaction, and the Planck scale (about 10^{16} TeV, or 10^{-35} m) at which gravity becomes unified with the other forces. The feebleness of gravity means that we require tremendous energies — far beyond what our accelerators will ever generate — if we wish to investigate the unification of all the forces.

The extreme weakness of gravity might have been the received wisdom, but nobody understood *why* gravity was so weak. How could we explain the factor of 10^{16} separating the two fundamental scales in nature? Quantum effects should drag the electroweak scale up to the Planck scale. This is the notorious hierarchy problem, a version of which we discussed on page 125. In many ways, finding a solution to the hierarchy problem has been one of the driving forces in modern high-energy physics. As we saw, the Standard Model does not even attempt to explain the disparity in scales; instead, the parameters of the model are adjusted so that the theory fits the observed electroweak scale.

The best solution to the hierarchy problem that theorists had come up with was to modify the Standard Model at the TeV scale. By introducing supersymmetry we stabilize the electroweak scale and keep it separate from the Planck scale. Notions of symmetry have been central in all our discussions to date, but what ADD suggested is that we can solve the hierarchy problem without resorting to symmetry. We can use geometry instead. Instead of supersymmetry modifying physics at the TeV scale, ADD suggested that it is the *Planck* scale that is modified by the presence of extra dimensions. If Newton's law is changed by the presence of large extra dimensions, then gravity could reach Planck strengths at distances much larger than the Planck length of 10^{-35} meters. In fact, if there are enough dimensions of large enough size it is possible to bring the Planck scale all the way down to the electroweak TeV scale. The hierarchy problem simply vanishes! In this picture there is only *one* fundamental scale in nature: the electroweak scale at 1 TeV.

For a given number of extra dimensions, one can calculate how large they must be if gravity is to be as strong as the other forces at the TeV scale. If there is one extra dimension, then its radius must be about 100 million kilometers in size. We can immediately exclude this possibility by observation, since 10^8 kilometers is roughly the distance of Earth from the Sun; over such distances we are certain that gravity obeys an inverse square law rather than an inverse cube law required by one extra dimension. For two extra dimensions, their size is about 1 millimeter. But, when ADD published their idea in 1998, gravity had not been tested on such scales. So it was entirely possible that gravity followed an inverse fourth power law

on submillimeter distance scales. ADD could solve the hierarchy problem without resorting to supersymmetry, so long as the Standard Model particles were restricted to a 3-brane and gravity alone was free to explore the two extra millimeter-sized dimensions. For a larger number of dimensions, the size required to resolve the hierarchy problem became smaller. For the maximum number of extra spatial dimensions allowed by M-theory — seven — the radius of these dimensions was 10^{-14} meters. This may seem small, but it is still very much larger than distances commonly occurring in particle physics; it is actually the size of a large atomic nucleus.

Although the ADD idea appears not to violate known physics, how does it fit in with the speculative ideas of physics beyond the Standard Model? How does it chime with the ideas discussed in earlier chapters? For example, what about grand unification? We saw how theories of supersymmetric grand unification provide convergence of coupling constants — and a grand unified theory of the gauge forces — at 10^{16} GeV. Although supersymmetric grand unified theories have not yet been positively demonstrated in experiment, they have several theoretical successes to their credit, such as the correct prediction of one of the key parameters of the Standard Model. Furthermore, the unification scale is high enough to explain why we do not see proton decay. But if quantum gravity is the appropriate theory at 1 TeV then presumably a grand unification of the forces does not occur. Do we have to lose all the benefits of grand unification in the ADD picture?

Well, soon after the ADD paper was published, three physicists then based at CERN — Keith Dienes, Emilian Dudas, and Tony Gherghetta — showed that if the gauge forces can explore *some* of the extra dimensions (but not the large dimensions that are the preserve of gravity alone), grand unification can occur at the TeV scale. Accurate calculations are difficult in their theory, however; and the fact that grand unification seems to occur in ordinary quantum field theory in four dimensions must be regarded as a fluke. Ignatios Antoniadis, working with Constantin Bachas of the Ecole Normale Superieure, developed a different approach that depends upon the properties of the bulk. Their approach seems to allow grand unification at the TeV scale and at the same time explains why the standard approach is so successful. Although there is much work to be done in this area, it is at least possible that grand unification occurs within the ADD scenario.

The idea of large extra dimensions is incredibly fertile. We have seen how it accounts for the hierarchy problem. The boxed text on page 256 discusses another example in which large extra dimensions could play a role: the problem of dark matter. In fact, with a little ingenuity, large extra dimensions can be brought to bear on several outstanding problems.

Large Extra Dimensions and Missing Matter

One of the mysteries of modern cosmology is the observation that much of the mass of the Universe must be in the form of dark matter. We do not see it, but we know it is there because of its gravitational influence on visible material. (The velocities of galaxies in clusters are too large to be explained by the gravitational effect of the visible mass inside the clusters. More convincingly, rotation curves — a rotation curve is the way in which rotational speed varies with distance from the center of a spiral galaxy — strongly indicate the presence of non-visible mass in galaxies.)

There are many candidates for particles that might comprise the dark matter of the Universe. We saw on page 131 that extensions to the Standard Model might provide a solution; for example, dark matter could be in the form of LSPs, the lightest supersymmetric particle. So although there has been speculation, no one knows for sure what dark matter consists of.

The ADD idea provides a quite different solution to the problem: dark matter could be ordinary matter that resides on a different brane, a parallel universe that is, say, 0.1 millimeter away from our own Universe. These are not parallel universes in the sense depicted in many science fiction novels; they are not universes identical to our own except that Hitler was victorious in World War II or John F. Kennedy was never assassinated. (Although if there were an infinity of parallel universes, perhaps even these outlandish possibilities would have to be considered.) Rather, they are universes literally parallel to our own, like the pages of this book are parallel to each other, but with their own version of particle physics. They might possess a Standard Model of particle physics like our own, but equally they might possess different forces, even different numbers of dimensions. Whatever local physics they obey, any matter in those universes would be dark: photons, or the equivalent of photons in those parallel universes, would be confined to those universes. Nevertheless, we would be aware of "parallel matter" through its gravitational interaction with our Universe. Gravity, remember, can travel through the bulk between branes.

Another possible explanation of dark matter is that our Universe is folded in the extra dimensions. A galaxy could be a fraction of a millimeter away from us in the extra dimensions, yet billions of light years distant in the dimensions through which light can travel. If a galaxy was, say, 10^{16} light years distant then its light would not have had time to reach us since the Big Bang. Yet it could affect us gravitationally, since it is less than a millimeter away in the extra dimension. This would be "ordinary" matter interacting gravitationally with the clusters of galaxies we see; but it would appear to us to be dark matter, since its light has not yet had time to reach us.

Probing the Higher Dimensions

There is a fifth dimension beyond that which is known to man. . .
This is the dimension of imagination.
Rod Serling, "The Twilight Zone"

The most intriguing aspect of the ADD proposal is that unification takes place at the TeV scale. If ADD are correct, we can test ideas that were once thought to be relevant only in the inaccessible Planck region of 10^{16} TeV. There are at least three ways in which we can hope to test these ideas: from astrophysical observations, from tabletop gravity experiments, and in particle accelerators.

Astrophysical tests of large extra dimensions

We have already seen how astronomical observations can put a loose limit on the number of large extra dimensions. If we want to solve the hierarchy problem with extra dimensions, observations of the Solar System tell us there must be more than one such dimension. (If there were only one large extra dimension, it would have to be about 10^8 kilometers in size, which is ruled out because planetary orbits obey Newton's inverse-square law.) So these observations tell us there must be at least two extra dimensions; observations and theoretical modelling of type II supernovas (see the boxed text on page 258) suggest there must be at least three extra dimensions.

Astronomers now have computer models of what happens during a type II supernova event, and it is clear that neutrinos carry off most of the energy of the explosion. A supernova emits *vast* numbers of neutrinos — about 10^{57} in a typical event. And this is a problem for the ADD idea.

The problem arises because, in ADD models, gravitons interact more strongly with matter than in traditional theories; they do so because gravity is stronger at shorter distances. In turn, this means that high-energy particle collisions in ADD models produce more gravitons than in traditional theories. Once produced, however, the gravitons are free to move in all the dimensions of space; they can remove energy from our brane and transfer it into the bulk. The problem? Well, if supernovas radiate too much energy into the bulk, they should sputter like damp firecrackers; but supernovas are sometimes bright enough for us to see them from halfway across the Universe. We *know* how most of the energy is carried away in a type II supernova: via neutrinos. Very little of the energy can be carried away by gravitons into the bulk. So our observations of supernovas set limits on how strongly gravitons can couple to matter. For two extra dimensions, it turns out that the Planck scale cannot be less than about 50 TeV.

Type II Supernovas

The supernova SN1987A, which occurred in the Large Magellanic Cloud, was about 170,000 light years from Earth. That made it close enough for astronomers to observe it in detail. A type II supernova like SN1987A occurs when a massive star exhausts its supply of nuclear fuel. A high-mass star rather quickly exhausts its supply of hydrogen, and then it starts shining through a series of nuclear reactions involving progressively heavier elements: carbon and oxygen nuclei fuse into neon and magnesium nuclei; neon and magnesium nuclei fuse into silicon and sulfur nuclei; silicon and sulfur nuclei fuse into iron nuclei. Towards the end of its life, the interior of a high-mass star resembles the layered structure of an onion. The star is doomed when it begins to develop an iron core. Any nuclear reaction involving iron *absorbs* energy, so the fuel supply switches off. If the core no longer produces energy, it no longer supports the star against the relentless force of its own gravity. When the core consists mostly of iron, a catastrophe occurs: the core begins to collapse under its own weight. In about a tenth of a second the inner core collapses into what is essentially a large neutron, and we have a neutron star. Vast numbers of neutrinos are created in this process. The contraction of the inner core is so quick that it "overshoots" and momentarily becomes even more dense than an atomic nucleus. At this point, the nuclear force stiffens the core and (if the core is not too massive) prevents further collapse. The inner core bounces back, like a rubber ball springs back if you squash it too tightly in your hands. Meanwhile, the outer core is still falling inwards. The rebounding inner core smashes into the infalling outer layers, and an explosion takes place. Between 0.3–0.5 of the star's mass is ejected into space; the remnant is a neutron star (or a black hole, if the core mass was large enough).

FIGURE 129 The blue supergiant star Sanduleak −69° 202 (arrowed in the right-hand panel) spent most of its life with a mass of between 15–18 times that of the Sun. It underwent a supernova explosion, which was detected by telescopes on Earth on February 23, 1989. (The same star field is shown in the left-hand panel; notice how this single star outshines the rest of the stars in this field put together.) The star became famous in the astronomical community under its new name, SN1987A.

If we want to use the ADD idea to solve the hierarchy problem, and bring the Planck scale down to about 1 TeV, then there must be at least three extra dimensions. (The more dimensions there are, the less strongly gravitons couple to matter because the interaction takes place at much shorter distance scales. If there were four or more extra dimensions, the ADD idea would pass any constraints coming from supernova observations.)

Cosmology provides constraints, too. When we look deep into space, we are looking at a time when the Universe was much younger than it is today. Cosmologists now have a successful model of how the Universe has evolved, from a fraction of a second after the Big Bang when the Universe was tiny and hot, to its present large, cool state. In the Standard Model of cosmology, the Universe cooled because it expanded, and this picture is in agreement with many observations. If the Universe is a brane, however, it cools because it is expanding and because it radiates gravitons into the bulk. Just as a hot coal radiates photons that carry off energy and cool the coal, so would our brane radiate gravitons that would carry off energy and cool the Universe. But the Standard Model of cosmology tells us that this could not have happened to an appreciable extent. The radiation of gravitons into the bulk causes another problem: some of these gravitons would return to the brane, where they could decay into extremely energetic photons. Astronomers have observed no sign of such events.

A very different test of the ADD idea comes from cosmic ray studies. Every second, about a hundred cosmic ray particles with energies of the order of a GeV strike every square meter of Earth's atmosphere; such energies are not particularly high. But once a week, on average, a cosmic ray particle with an energy of 10^6 TeV hits an area of a square kilometer. Those are really high-energy events, surpassing what we generate in particle accelerator experiments. About once a century, a cosmic ray particle with an energy of 10^8 TeV hits an area of a square kilometer. This energy is *way* beyond what we can generate in experiments. The origin of these ultra-high-energy particles is a great mystery, since they must originate fairly close to us. (Their origin must lie within about 180 million light years of Earth; had they travelled a greater distance, collisions with the cosmic microwave background would have reduced their energy. Now 180 million light years seems a huge distance, but it is in the near neighborhood on a cosmological scale.) However, there are no nearby astronomical objects that could generate such high-energy particles.

Even the highest-energy cosmic rays cannot reach the Planck scale, at least not within traditional theories. As we have seen, however, if large extra dimensions exist then less energy is required to reach the Planck scale. The following question then arises: might the collision of a cosmic ray with the atmosphere have energy enough to probe Planck-scale physics?

A New Generation of Cosmic Ray Detectors

Partly to solve the origin of ultra-high-energy cosmic rays, astrophysicists are developing new cosmic-ray observatories.

When a cosmic ray particle hits the atmosphere with an energy of 10^8 TeV it creates a shower of secondary particles that may cover an area as large as 16 square kilometers by the time it reaches Earth's surface. Therefore, in order to fully analyze such large showers, astrophysicists require an observatory with a large area.

The Pierre Auger Cosmic Ray Observatory in Argentina consists of 30 fluorescence telescopes overlooking 1600 particle detector stations arranged in a grid that covers 3000 square kilometers — more than three-and-a-half times the area of New York City. A similar observatory, covering the Northern hemisphere, will eventually be built in Utah. So far, the project includes more than 250 scientists from 16 countries.

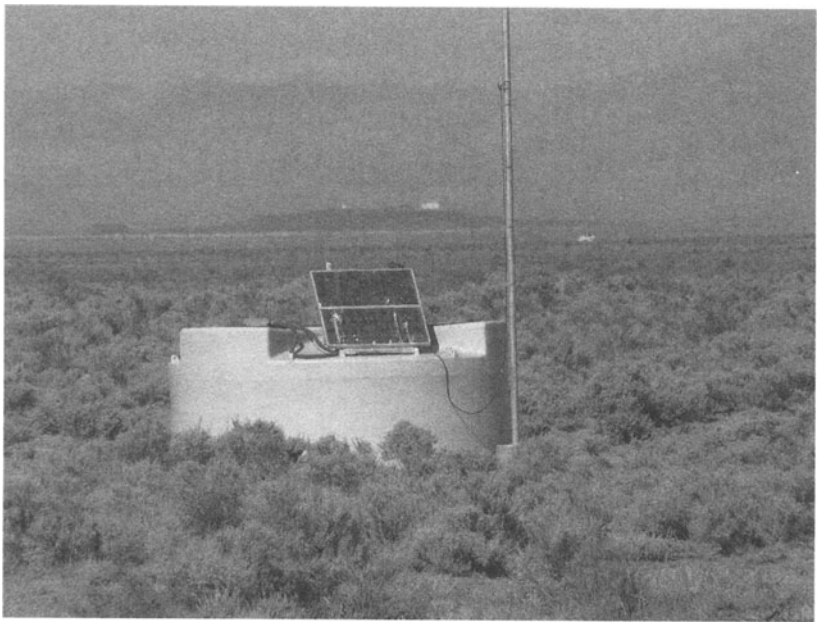


FIGURE 130 One of the cosmic ray detectors that form part of the Pierre Auger Cosmic Ray Observatory. These detectors are 3000-gallon water tanks; inside, the tanks are totally dark — except when a particle from a cosmic-ray shower passes through. The particle produces light in the form of Cerenkov radiation, which can be detected and measured; many of the properties of the original particle can then be deduced. There are 1600 of these detectors, spaced at 1.5 kilometer intervals, over 3000 square kilometers of a plain called Pampa Amarilla in Argentina. A similar observatory will be constructed in Utah. Together, the detectors will provide information on the ultra-high-energy cosmic rays that initiate the showers.

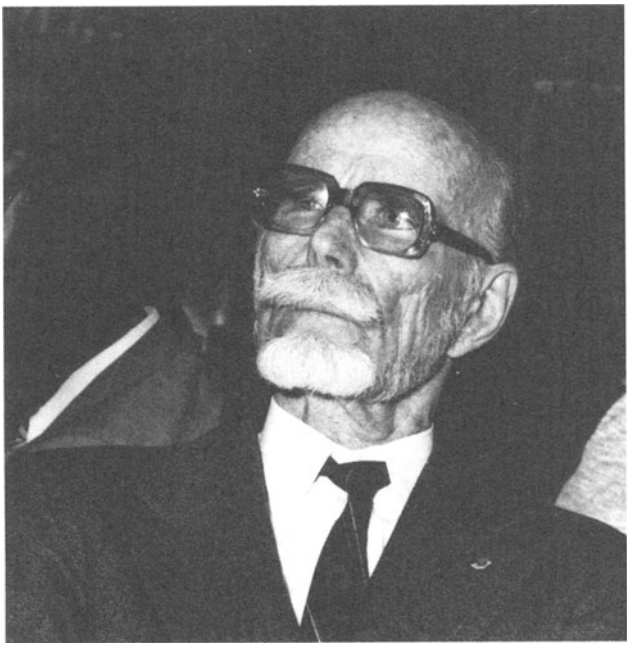


FIGURE 131 In 1938, Pierre Auger discovered "extensive air showers" — showers of secondary particles caused by the collision of high-energy primary cosmic rays with the atmosphere. He thus deduced that cosmic-ray particles could possess an energy of at least 10^6 GeV, which was a million times more energy than particle accelerators could produce. Although the energy available in particle accelerators has increased over the years, so has the maximum energy detected in cosmic-ray particles. The highest-energy cosmic rays are still much more energetic than any particles created in our laboratories. Auger was convinced that the study of ultra-high energy cosmic rays was important for physics — though even he could not have imagined that such a study might give us glimpses of extra dimensions! — so it is appropriate that the Pierre Auger Cosmic Ray Observatory is named after him.

Well, in 2001, one of the ADD authors — Savas Dimopoulos — teamed up with Greg Landsberg, a physicist at Brown University, to show that if large extra dimensions exist, high-energy collider experiments at the Large Hadron Collider might be energetic enough to produce microscopic black holes. At about the same time Steven Giddings and Scott Thomas, two physicists collaborating at the Santa Barbara Kavli Institute for Theoretical Physics, reached a similar conclusion. A few months later, Jonathan Feng of the University of California, Irvine and Alfred Shapere of Kentucky University pointed out that if the Large Hadron Collider can produce microscopic black holes, some cosmic rays will *certainly* possess enough energy to produce black holes. In fact, in the time it has taken you to read this paragraph, a black hole should have been created somewhere in Earth's atmosphere.

Extra Dimensions and the Highest Energy Cosmic Rays

The ADD idea allows the imagination of physicists to run wild. They can now propose seemingly crazy solutions to difficult problems. For example, we have already mentioned the mystery of the highest-energy cosmic rays. The mystery is that these particles must have originated within 180 million light years of Earth. But it is extremely difficult to *accelerate* a cosmic-ray particle to such high energies, and even if one can think of mechanisms to do it, there seems to be no astronomical object within 180 million light years of us capable of providing such an acceleration. The ADD picture allows for a quite different explanation of these ultra-high energy cosmic rays. Suppose a black hole is produced somewhere in a cosmic-ray interaction and then it moves away into the bulk. (Nearly all black holes will remain on the brane, since they are likely to possess an electric charge, or a color charge, or a lepton number — Standard Model properties that are necessarily confined to the brane. But there are scenarios where black holes can wander off into the bulk; for example, a graviton could be created on the brane and then move into the bulk where it becomes a black hole.) In some models the black hole remains close to the brane, where it accretes mass and slowly increases in size. Once it reaches the Planck mass (10^{16} TeV), the black hole event horizon becomes so large that it touches the brane; the hole then evaporates onto the brane, typically into about ten particles each with energy 10^{15} TeV. In this picture, these black-hole remnants *decelerate* to about 10^{10} TeV (which is much easier than *accelerating* a particle to about 10^{10} TeV) and later become the ultra-high energy cosmic rays that we puzzle over. This is an entirely speculative notion — but it is an interesting way of looking at the phenomenon.

Any such black hole would quickly decay into a shower of Hawking radiation (mainly into Standard Model particles on our brane, rather than into gravitons into the bulk). This shower of radiation would be quite different from showers arising from, say, the collision of a cosmic-ray proton with an atmospheric atomic nucleus. Gravity is "flavor blind," so when a microscopic black hole evaporates it produces all the Standard Model particles with essentially equal probability. Once one accounts for spin and color, it turns out that the particles produced when a black hole decays are about 72 percent quarks and gluons, 18 percent leptons, and the rest are bosons. Such a distinctive shower of particles would be hard to miss. So there is a possibility that the Pierre Auger Observatory will detect black holes. We saw in an earlier chapter how cosmic-ray experiments played a

key role in the development of particle physics; their importance lessened with improvements in colliders, but cosmic-ray observations have now regained their importance. Observatories such as Auger may well give us our first glimpse of particle physics beyond the Standard Model. (Incidentally, we need not worry about the possible creation of microscopic black holes: they cannot harm us. The black holes are so small they have only a fleeting existence. They decay into Hawking radiation after about 10^{-26} seconds.)

Different astrophysical observations can test other aspects of the ADD idea. For example, a new generation of gravitational-wave detectors will soon come on line that will have the capability of detecting gravitational waves emanating from violent astrophysical events (see the boxed text on page 264). These detectors might also soon test the ADD-inspired possibility that dark matter is due to "ordinary" matter — stars, galaxies, quasars — that are on a different fold of our Universe. Suppose that a violent event takes place in a neighboring fold: the collision of two neutron stars, for example. Such an event would be an intense source of gravitational waves, which would reach us quickly via the shortcut of extra dimensions and be detectable by LIGO (Laser Interferometer Gravitational Wave Observatory) and similar detectors. Conventional telescopes, however, would detect no electromagnetic radiation from the collision, since it might take radiation billions of years to reach us along the fold. So sources of gravitational radiation with no visible counterpart could be a signal of extra dimensions.

Tabletop gravity tests of large extra dimensions

After the speculative reaches of astrophysics, it can be a relief to return to the solid reality of an experiment that can fit on the top of a table. It is also a quite wonderful thought that such experiments — tabletop gravity experiments — might provide our first glimpse of extra dimensions.

When ADD proposed the idea of large gravity-only dimensions, gravity had only been tested down to 1 millimeter. The classic way to measure gravity on such scales is to use a torsion balance. This is what Henry Cavendish used in 1798 to measure Newton's gravitational constant, and variants on the apparatus are used to this day. The Cavendish setup consisted of a light rod suspended at its center by a thin wire. At each end of the rod was a small lead ball. The rod was free to twist about the wire. The setup was so sensitive that the minuscule gravitational force produced by two large lead balls placed either side of the small lead balls was enough to produce a measurable twist. Unfortunately, other types of force can cause the rod to turn: static electrical charges, vibrations in the floor, air currents, and so on. The difficulty of performing this experiment is the reason why, by 1998, the inverse-square law had only been tested down to 1 millimeter.

A New Generation of Gravitational Wave Detectors

The American observatory called LIGO — Laser Interferometer Gravitational Wave Observatory — has recently begun operation. Similar observatories include VIRGO (an Italian–French collaboration) and GEO (a German–British collaboration). The Japanese TAMA is well advanced, and there are several other observatories in various stages of development. Soon, Earth will be girdled with gravitational-wave detectors, and a new window will open up on the Universe.

Such detectors will be able to directly verify that gravitational waves exist and will test the prediction that the graviton is a spin-2 particle with zero rest mass. They will be able to confirm that black holes exist and test general relativistic predictions for what happens when two black holes collide. They will also teach us much about the nature of supernovas, neutron stars, and galactic nuclei. They may also be able to detect the ripples in spacetime that would have generated when space and time came into being in the Big Bang.

A further exciting possibility — though it might be long-shot — is that an analysis of gravitational radiation will provide evidence for the existence of higher dimensions.

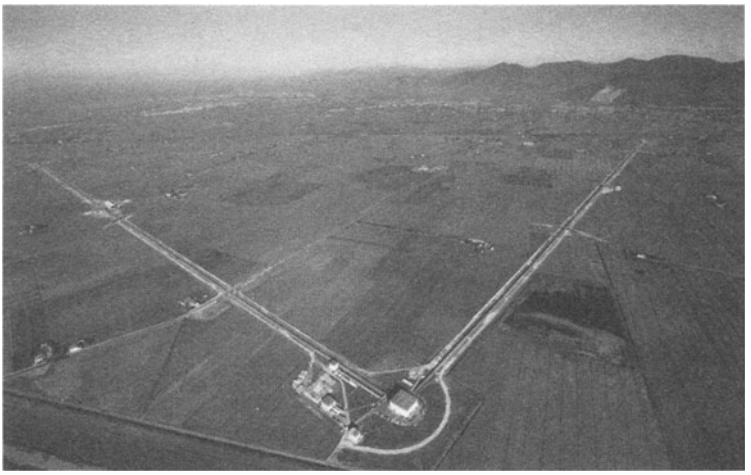


FIGURE 132 The VIRGO interferometer lies a few kilometers from the town of Pisa in Italy, where Galileo performed his experiments with gravity. The two 3-kilometer-long wings of the interferometer are clearly seen in this aerial view. VIRGO is an Italian–French collaboration; GEO, which is located near Hannover, is a German–British collaboration with a similar design to VIRGO. Other prominent gravity wave detectors include the LIGO interferometers in the U.S., and TAMA in Japan. Still other detectors are in various stages of data-taking or planning. Gravity wave research will provide us with a completely new window on the Universe.

Once ADD had raised the possibility of probing extra dimensions by measuring deviations from the inverse-square law at the millimeter level, physicists immediately began making plans for doing exactly that.

In 2001, for example, a group led by Eric Adelberger and Blayne Heckel at the University of Washington used a vastly improved version of the torsion balance to measure the strength of gravity down to shorter ranges than had, until recently, been thought possible. Their experiment (see Figure 133) found that Newton's law of gravity holds down to 0.2 millimeters. So the extra ADD dimensions, if they exist, can be no larger than 0.2 millimeters.

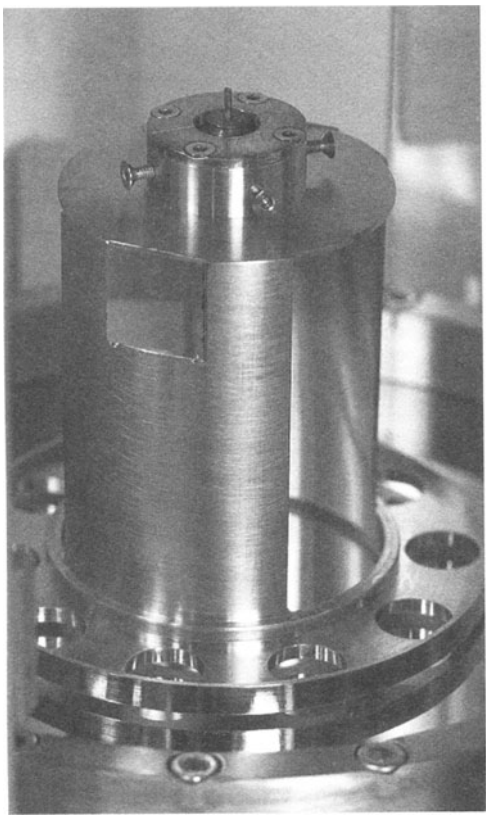


FIGURE 133 This apparatus — known as a torsion pendulum — has been used to search for extra dimensions. The pendulum is suspended by a 20-micron-thick tungsten fiber (barely visible in this photograph) and is free to rotate. Any rotation is measured by shining a laser beam off a square mirror that is attached to the upper part of the pendulum. The active component of the pendulum is a ring with ten equally spaced holes bored into it. Directly below the ring is a rotating attractor consisting of two disks, each of which also has ten equally spaced holes. The holes in the lower, thicker disk are offset by 18° compared to the upper disk: the holes in the lower disk thus lie halfway between the holes in the upper disk. If gravity is stronger than we believe at the millimeter scale, the twist on the ring induced by the lower disk will be too small to cancel the twist induced by the upper disk. As the attractor rotates there will be a force acting on the pendulum that varies ten times for each complete rotation of the attractor. But if the inverse square law is correct at these scales, the twist produced by the two disks exactly cancel.

Other groups have tried rather different approaches to the measurement of gravity on small distance scales. For example a group led by John Price, of the University of Colorado at Boulder, measured the gravitational attraction between two sheets of tungsten that are just 20 millimeters long and 0.3 millimeters thick. These wafer-thin sheets were placed 0.108 millimeters apart; one of the sheets, which acted as the "source mass," was

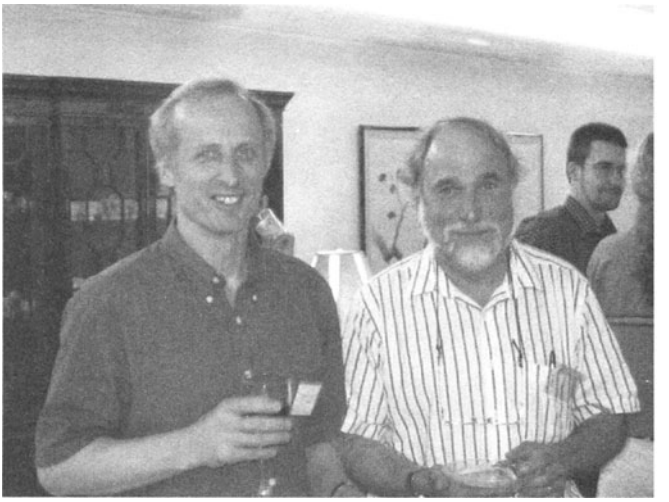


FIGURE 134
Blayne Heckel (left) and Eric Adelberger (right) have shown that Newton's inverse square law of gravity holds down to 0.2 millimeters. If extra dimensions exist, they must be smaller than 0.2 millimeters.

set into vibration at the resonant frequency of the second sheet, the "test mass." The size of the vibrations in the test mass enabled Price to calculate the gravitational attraction between the two tungsten sheets — and it was exactly what Newton would have calculated. As of the time of writing, this result by Price's group is the shortest distance over which gravity has been measured, and there are no signs of extra dimensions. (Interestingly, the results are also a test of certain string theories. Some models based on string theory suggest that the dilaton should modify Newton's law on distances of order 0.1 millimeters; Price's result rules out these models. Some aspects of string models are becoming testable!)

Particle collider tests of large extra dimensions

When the idea of large extra dimensions was proposed, the experimental high-energy physics community moved quickly to test it. Physicists combed through their results at particle accelerators like LEP (which, until its dismantling, was used to collide electrons with antielectrons) and Tevatron (which collides protons with antiprotons) looking for any signatures of higher dimensions that they might have missed.

The key for the experimenters is that the ADD dimensions, if they exist, bring certain quantum gravity effects within reach. In particular, at the energies available at LEP and Tevatron, particles become so close to each other during collisions that they might produce gravitons. There are two ways in which gravitons could make their presence felt: by graviton emission into the bulk and by graviton exchange processes interfering with Standard Model results.

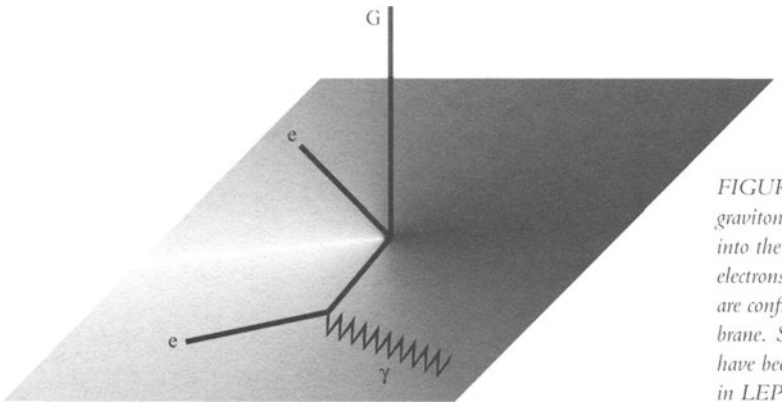


FIGURE 135 A graviton (G) escaping into the bulk. The electrons and photon are confined to the brane. Such effects have been looked for in LEP data.

Consider the process shown in Figure 135, a new form of Feynman diagram, in which an electron–antielectron pair interact to produce a photon and a graviton. Such a process might have taken place at LEP. The photon, because it is a Standard Model particle, is confined to our brane. But the graviton is free to roam the higher dimensions, carrying away energy from our brane. We would see an electron and an antielectron enter, and a photon leave; once we calculated the total energy before and after the interaction we would notice that some energy had gone missing. Something similar could occur at the Tevatron, this time involving interactions between the quarks inside protons rather than electrons. (Imagine a diagram similar to that in Figure 135, but with the "e" replaced by "q.") One can imagine a proton–antiproton pair colliding and producing jets of gravitons. Again, collisions in which energy simply vanishes could be a signal of gravitons carrying energy into the bulk.

Of course, this is easier said than done. Experimentally, it is extremely tricky to look for things that are missing. Neutrinos, for example, can leave the scene of a collision without being detected. But the ADD models predict how graviton production will vary with increasing energy, so if a signal is there to be found then experimenters will find it. So far, researchers have combed through the data generated by Run 1 at Tevatron (data taken between October 1992 and February 1996) and found no indication of extra dimensions. Similarly, researchers have analyzed the results from LEP up until it stopped taking data in December 2000, and again they found no sign of extra dimensions. This is ongoing work, and there may yet be surprises hidden in the data, but at the time of writing there has been no sign of graviton emission.

The second way that Tevatron and LEP might produce evidence of extra dimensions is indirectly, through graviton exchange processes that alter the predictions of the Standard Model. For example, when an electron and

antielectron come close together they interact via the exchange of a virtual photon. This is the standard electromagnetic interaction. In the ADD scenario, they can also interact via the exchange of a virtual graviton. This process gives rise to a different type of Feynman diagram; it is a process that until recently had not been worth considering when calculating the details of electron interactions. Again, experimenters have trawled extensively through their data, looking for telltale deviations from the Standard Model that might indicate the presence of extra dimensions. And again, at the time of writing, nothing has been found.

We need not be discouraged, however. In 2007, if everything goes to plan, the Large Hadron Collider will commence its experiments. The LHC will fire two proton beams into each other at speeds just fractionally short of the speed of light. Each beam will have an energy of about 7 TeV; the energy involved in head-on collisions will be much larger than those available to the Tevatron, which is currently the most powerful particle accelerator on the planet. The energies available at the LHC should certainly be high enough to reach into the extra dimensions — provided that the ADD idea is correct.

The extra power of the LHC opens up other possibilities for testing the ADD idea beyond looking for indirect signals of graviton emission. One of the possible outcomes of the LHC experiments is the production of black holes by the bucket-load. As mentioned in our discussion of cosmic rays, Landsberg and Dimopoulos, and independently Giddings and Thomas, pointed out that if large extra dimensions exist, the LHC could become a black hole factory. It could make a black hole every second. The cathedral-sized detectors of the LHC would easily pick up the burst of Hawking radiation from the near-instantaneous decay of the black holes. This is all rather different than what physicists anticipated when they designed the machine: they thought they would be probing smaller and smaller distances, not extra dimensions.

Another sign of large extra dimensions that might show up at the LHC would be excitations of strings. As we know, in traditional string theories the familiar particles — electrons, quarks, photons, and so on — are the fundamental vibrational modes of strings. The strings have usually been thought of as being 10^{-35} meters long, but if the ADD scenario holds true then strings could be much longer — about 10^{-19} meters long — and so the LHC might be able to "pluck" the strings into quite different vibrational modes. If this happened, then we would see a host of exotic particles — such as high-mass versions of the familiar electrons, quarks, and photons. The prospect is enticing.

There is yet another possibility — and no doubt there are other signs of extra dimensions that physicists have not yet thought about. If we do live

on a brane, and if there are extra dimensions, then the brane can fluctuate along those extra dimensions. The rules of quantum mechanics demand that such fluctuations manifest themselves as quantum particles — *branons*. These new additions to the bestiary of hypothetical particles would be scalar (spin-0) particles. Their mass would depend on the manner in which the extra dimensions are curled up. (If the extra dimensions are flat, the branons will be massless.) Physicists have looked for the interaction of branons with Standard Model particles in LEP data, so far without success. Again, the increased energies available to the LHC may give it the capacity to generate branons — so experimenters have something else to look for.

If the LHC finds no telltale signals from gravitons, if it produces neither black holes nor string excitations nor branons, it will at least have provided tight constraints on the ADD idea — tight enough, perhaps, to rule it out completely.

Why Two Branes are Better Than One: The Randall–Sundrum Scenario

My Lord, your assertion is easily put to the test.
You say I have a Third Dimension, which you call "height."
Now, Dimension implies direction and measurement.
Do but measure my "height," or merely indicate to me
the direction in which my "height" extends, and I
will become your convert.
Edwin A. Abbott, *Flatland*

The ADD idea is a tremendously fertile one from the experimenter's point of view: for the first time in decades, it seems as if there might be a way to search for completely new fundamental physics. But from the theoretical point of view, the idea of large extra dimensions has a problem.

The prime motivation behind the ADD idea is that it claims to solve the hierarchy problem — the problem of why gravity is so weak; in other words, the troublesome fact that the Planck scale is 16 orders of magnitude greater than the electroweak scale. However, the "solution" is to say that an unspecified number of dimensions are many orders of magnitude larger than the natural compactification scale. As we have seen, if there is one large extra dimension then it must have a radius of 10^{11} meters in order to solve the hierarchy problem; if there are seven large extra dimensions, then they must have a radius of 10^{-14} meters in order to solve the hierarchy problem. Even the latter value is huge compared to the compactification

radius that is usually assumed: the Planck scale of 10^{-35} meters. While it would be nice for experimenters if these large extra dimensions existed and produced effects that we can observe at the TeV scale, there is no obvious reason why the compactified dimensions should be so much larger than previously supposed. In a sense, the ADD scenario replaces one hierarchy (that between the electroweak scale and the Planck scale) with another (that between the electroweak scale and the compactification scale). In one sense, it has not solved the hierarchy problem at all; it has merely reformulated the problem.

A few months after the appearance of the ADD paper, Lisa Randall, who was then at Princeton University, and Raman Sundrum, who was then at Boston University, developed a theory that solved the hierarchy problem in a much more elegant, beautiful way. Their theory required just one extra dimension, rather than the minimum of two extra dimensions of the ADD scenario, and in some formulations of their theory, there wasn't even a need for compactification: the extra dimension could be infinite in extent!

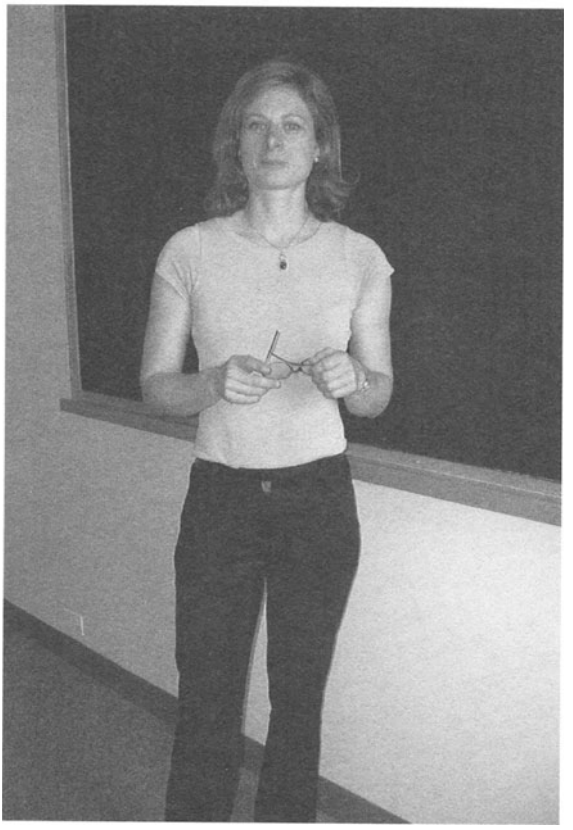


FIGURE 136 Lisa Randall is a professor of physics at Harvard University. In addition to her research on extra dimensions of space, she continues to research supersymmetry and other physics beyond the Standard Model.

Randall and Sundrum pointed out that a brane carries energy so, according to general relativity, a brane embedded in a bulk space will warp the bulk space around it. Suppose there is a 3-brane — called the *Planck brane* — embedded in a spacetime with one extra spatial dimension. For certain configurations, the Planck brane can warp the fifth dimension so that gravity stays close to the brane. In other words, there is a "warp factor" (not to be confused with the warp factor of "Star Trek") associated with the extra dimension. In essence, the warp factor gives the probability of finding a graviton at any particular point in the extra dimension. In the Randall–Sundrum theory, the warp factor is such that there is a large probability of finding a graviton close to the brane — so gravity is strong there. Indeed, gravity could be as strong as any of the other forces. But as we move away in the extra dimension, the probability of finding a graviton decreases exponentially — so gravity is exponentially weaker as you move away from the brane. Thus a particle that is on the brane would be heavy, since gravity is strong there; but if the particle could move away from the brane in the extra dimension it would rapidly become much lighter. Note that even if the extra dimension is infinite and even if gravity can in principle propagate throughout the entire dimension, in practice gravity stays localized to the brane. There is no need to compactify the extra dimension: it just *appears* as if the extra dimension of space is compactified because gravity is likely to venture only a small amount into that dimension.

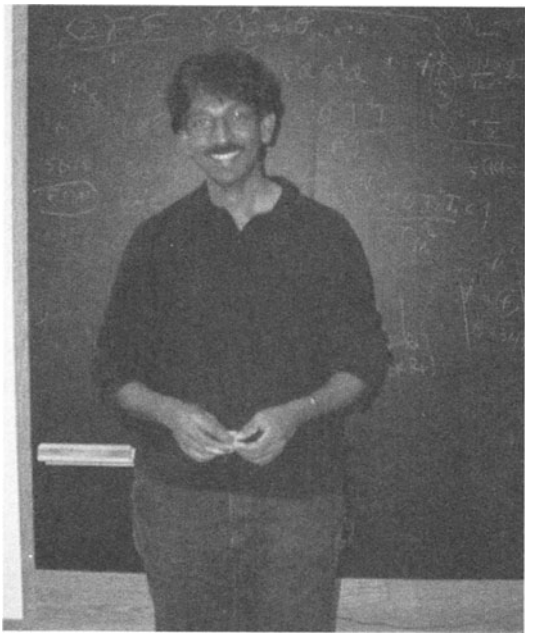


FIGURE 137 Raman Sundrum is a professor of physics at the Johns Hopkins University. He researches the observable implications of supersymmetry and extra spacetime dimensions.

Now add another 3-brane into the picture. This brane — called the *TeV brane* — is parallel to the original Planck brane and is just a small distance away in the extra, fifth dimension. This is the brane that traps the Standard Model particles to itself. This is the brane we live on. Gravity on the TeV brane is weak because the warp factor "dilutes" gravity from the Planck brane. In this scenario the hierarchy problem has nothing to do with wildly different energy scales or unaccountably large dimensions; it is simply caused by the relative location of two 3-branes in a five-dimensional spacetime. If our TeV brane were closer (in the extra dimension) to the Planck brane, we would measure a stronger gravitational interaction; if our TeV brane were farther away, gravity would be even weaker than we measure it to be. The hierarchy problem is solved here much more naturally than in other approaches. In the original formulation of the hierarchy problem, there was a ratio of 10^{16} between the two energy scales; in the Randall–Sundrum scenario, the "distance" between the two branes, in appropriate units, is of the order of ten. And a factor of ten is much easier to generate than 10^{16}.

The Randall–Sundrum theory involved a particular type of geometry that solved Einstein's equations. The 3-branes were embedded in a five-dimensional spacetime with a constant negative curvature — anti-de Sitter spacetime, AdS_5. This AdS_5 spacetime was interesting in its own right, of course, because it was central to Maldacena's work showing the correspondence between gravity and quantum field theories. The scenario is also similar to the Hořava–Witten spacetime, which arises in M-theory. Thus there appears to be a natural link between M-theory and the Randall–Sundrum scenario. Again it should be stressed that M-theory is not even defined yet as a theory. But it seems that the natural framework in which to discuss ideas like the Randall–Sundrum scenario may well turn out to be M-theory; certainly the ingredients of the scenario are found there.

If the Randall–Sundrum scenario is correct, are there experimental signatures we can look for? Tabletop gravity experiments will not find evidence for the scenario: gravity behaves exactly as Newton predicted it would. Although the Randall–Sundrum scenario is a five-dimensional theory, the warp factor ensures that it *looks* like a four-dimensional theory. So the tabletop gravity experiments should continue to observe an inverse square law for gravity. Things are different on the TeV scale, however. The brane we live on has a TeV scale associated with it. Once we get to TeV energies, physics should change and so experiments at particle accelerators could find evidence for the Randall–Sundrum scenario. At the energies available at the LHC, the Kaluza–Klein tower of particles associated with the five-dimensional gravitons should be produced. If the LHC were to find signs of a tower of spin-2 particles, we might interpret that as evidence

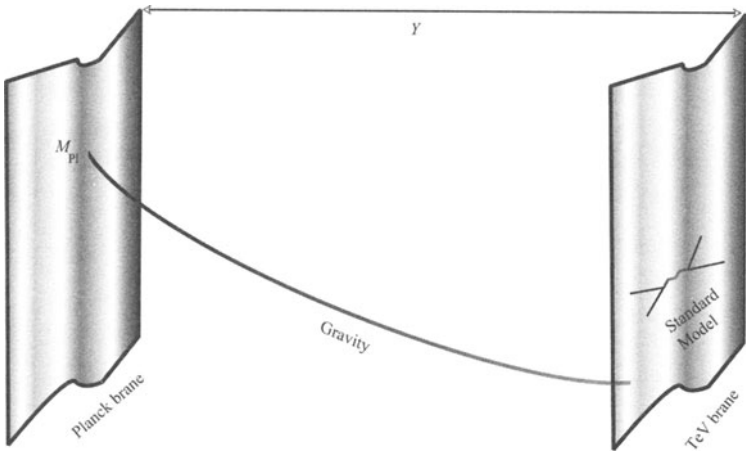


FIGURE 138 In the original Randall–Sundrum scenario, two branes (called the Planck brane and the TeV brane) are separated by a distance Y in a warped fifth dimension. Although gravity is free to roam through the bulk, the warped geometry ensures that gravity is localized to the Planck brane. The strength of the gravitational interaction falls off exponentially with distance from the brane. In this scenario, the strength of gravity on the TeV brane, where the Standard Model lies, depends simply on the distance Y.

in favor of the RS scenario. In the ADD scenario, the spacing between Kaluza–Klein energy levels is equal; it is also so small that the spectrum resembles a continuum. The RS scenario also has Kaluza–Klein states, but because of the warped geometry of the fifth dimension the spacing between the energy levels is no longer equal. Individual particles should be observable. Furthermore, whereas the ADD scenario has graviton excitations that are rather light, the RS scenario gives a mass of about 1 TeV for the first graviton excitation. This is heavy enough so that astrophysical observations do not constrain the theory, as with ADD, but it is still light enough so that the LHC should find them.

This is an optimist's game, and for the optimist there is another possible experimental signature of the RS scenario. In the original two-brane scenario, the size of the extra dimension could take any value and the position of the second brane was not determined by the model. Soon after the publication of the RS paper two physicists at the California Institute of Technology, Walter Goldberger and Mark Wise, studied ways of stabilizing the position of the second brane. They showed that a scalar field — called the radion field — would effectively determine the size of the extra dimension. The radion field would have to be a bulk field, and like the gravitational field it would have to propagate in *all* the spatial dimensions. But like gravity, it should also have observable effects in our four-dimensional world. The quantum of the radion field is called the *radion*. The radion is a mas-

sive spin-0 field, and its properties have much in common with the Higgs boson (in addition to the fact that there is no experimental evidence for either of these hypothetical particles). If experiments at the LHC do find the radion, we will have evidence for the existence of an extra dimension.

Randall and Sundrum noticed something else about branes in a warped spacetime geometry. The two branes — Planck brane and TeV brane — were introduced to solve the hierarchy problem. Suppose, though, that the hierarchy problem is solved in some other manner. In that case, we do not need two branes to describe physics. We might live on a Planck brane that is embedded in a five-dimensional spacetime in which the fifth dimension is infinite in extent, and we would not know it. In this scenario, because of the warp factor, events that happen far away from the Planck brane are essentially irrelevant to physics that happens on the Planck brane. A world with *four* infinite spatial dimensions would look very similar to a world with *three* infinite spatial dimensions. On page 143 I wrote that "If there is one thing we can be sure about from our everyday experience, it is this: the world possesses three large spatial dimensions." Now we know that our everyday experience could be quite compatible with there being a fourth, infinitely large dimension that is too warped for us to see. Ultimately, the true test of the dimensionality of space must be experiment, rather than prejudices arising from everyday life.

Randall and Sundrum thus presented two scenarios — called RS1 and RS2 — involving branes and warped geometry. (The numbers refer to chronology, not the number of branes. The first scenario, RS1, involves two branes; the second scenario, RS2, involves only one brane.) Their papers sparked a huge surge of interest, since both scenarios have something to say about realistic particle physics and cosmology. And, as with the ADD scenario, experiments may be able to find evidence in favor of them. In a few years time, at the LHC, experimenters may be able to tell us whether we are living on a brane.

Is Spacetime Really Symmetrical?

No amount of experimentation can ever prove me right.
A single experiment can prove me wrong.
Albert Einstein

The year 2005 marks the hundredth anniversary of Einstein's *annus mirabilis*. One of his three key achievements in 1905 was the introduction of Lorentz symmetry to describe the laws of physics. Lorentz symmetry — the demo-

cratic notion that the laws of physics are the same for all inertial observers — led Einstein to his theory of special relativity. We saw in Chapter 2 that one of the consequences of Lorentz symmetry is that the distinction between space and time is blurred: we live in a four-dimensional spacetime. But is Lorentz symmetry an *exact* symmetry of nature?

It turns out that M-theory allows for the possibility that Lorentz symmetry might not hold exactly. Tests for Lorentz violation might provide signals of new physics. They might be a probe of M-theory. Even if particle accelerators and tabletop gravity experiments fail to find evidence for M-theory, perhaps observations of Lorentz violation will do the trick.

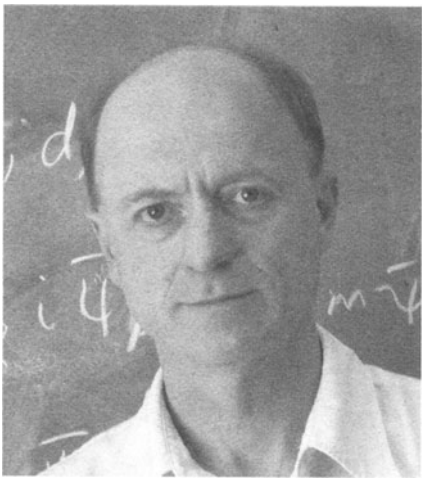


FIGURE 139 Alan Kostelecký

The study of violations of Lorentz symmetry has been pioneered by Alan Kostelecký, of Indiana University, and his coworkers. Kostelecký pointed out that there are two different formulations of Lorentz symmetry. *Observer* Lorentz symmetry is the statement that the laws of physics are the same for all inertial observers. *Particle* Lorentz symmetry is the statement that a particle having some velocity with respect to a fixed inertial frame experiences the same laws of physics as a particle at rest in the inertial frame. We usually consider these two formulations to be identical. For example, a passenger seated on a moving train and a person seated at a station platform experience the same laws of physics; this is an example of observer Lorentz symmetry. If the seated passenger gets up and walks along the carriage — if he undergoes a particle Lorentz transformation, in other words — what happens? We assume that, from the perspective of the moving train, a seated passenger and a moving passenger experience the same laws of physics; the moving passenger has merely introduced another equally valid frame of reference into the discussion. So we assume that the observer Lorentz transformation and the particle Lorentz transformation are the same. But Kostelecký showed that if Lorentz symmetry is not exact, the two transformations are not equivalent. Observer Lorentz symmetry can still hold while particle Lorentz symmetry is broken; the moving passenger can experience different laws of physics from those experienced by the seated passenger.

In 1989, Kostelecký and Stuart Samuel, a physicist at the City University of New York, showed that string theory possesses an interesting feature.

It seems to permit the spontaneous breaking of Lorentz symmetry in the early Universe. (We saw the importance of spontaneous symmetry breaking in Chapter 4, when we discussed electroweak unification.) If such a spontaneous breaking of Lorentz symmetry did occur, the Universe would now be filled with background fields that point in spontaneously chosen directions (just as magnetic dipoles align in some spontaneously chosen direction when they cool to form a magnet). These background fields would violate particle Lorentz symmetry but maintain observer Lorentz symmetry. An elementary particle moving in the presence of one of those background fields would undergo interactions that have a preferred direction in spacetime. Two identical experiments could, in principle, give different results merely because the two laboratories were pointing in different directions. The same experiment could also give different results at different times of the year, as the Earth moved in its orbit around the Sun.

The breaking of Lorentz symmetry in Kostelecký's theory would occur at the Planck scale of 10^{19} GeV, the region often claimed to be forever beyond the bounds of experiment. But Kostelecký and his coworkers have shown that effects of Lorentz symmetry violation *can* be detected — not by the crude means of colliding particles at high energies, but by making ultra-precise measurements of particles at low energies. How precise do we have to be? We cannot know for sure, since M-theory is not yet properly defined, but we can make an informed guess. The rest mass of the proton is roughly 1 GeV and the Planck mass is roughly 10^{19} GeV, so if we do experiments on protons that are sensitive to effects below one part in 10^{19} we would effectively be probing the Planck scale. And any violation of Lorentz symmetry would immediately jump out: nothing in the Standard Model permits the breaking of this fundamental symmetry.

There are many possible experimental tests of Lorentz symmetry. The classic test is the Michelson–Morley experiment, first conducted in 1887, the results of which helped inspire the theory of special relativity. The Michelson–Morley experiment attempted to detect a dependence of the speed of light on direction in space (it actually found no such dependence) and was thus a test of Lorentz symmetry. A related experiment first performed in 1932, the Kennedy–Thorndike experiment, tested Lorentz symmetry by looking for a dependence of the speed of light on velocity. In 2003, an updated version of these experiments — involving the monitoring of microwaves inside a pair of cylindrical cavities oriented at right angles to each other — was performed by a group at Stanford. As the Earth rotated on its axis and orbited the Sun, the group looked for changes in the resonant frequency of the cavity. None was found. Similar experiments involving the year-long observation of the resonant frequencies of resonators made of crystalline sapphire were performed, independently, by groups in

Berlin and Paris. They too found no sign of Lorentz violation. If there is a violation of Lorentz symmetry by the photon, these recent experiments show that it must be less than one part in 10^{11}.

Violations of Lorentz symmetry can also be searched for in matter. As Earth moves around the Sun the orientation of electron and quark spins might vary relative to the fixed background fields that arise from the spontaneous breaking of Lorentz symmetry. Variations in quark spin orientation, for example, would lead to variations in the frequency of masers (the microwave version of a laser). Experiments with masers can detect even tiny frequency changes, and such experiments have shown that violations of Lorentz symmetry by the neutron (which, as we know, consists of three quarks) must be less than one part in 10^{31}; for the proton, the limit is one part in 10^{27}. Variations in electron spin orientation can be measured with a version of the Adelberger–Heckel torsion pendulum. Using this apparatus, Adelberger and Heckel have shown that violations of Lorentz symmetry by the electron must be less than one part in 10^{29}.

There are other tests of Kostelecký's idea. For example, the symmetry of CPT is closely related to Lorentz symmetry. The CPT transformation consists of a combination of three transformations: swapping particles with antiparticles (an operation called charge conjugation, symbol C), reflection in space (P for parity), and reversing the direction of time (T for time reversal). Each of the transformations C, P, and T has been observed to be violated; so has the combination CP (see page 108). So far, however, the full combination of CPT appears to be an exact symmetry of nature. In 2002 Oscar Greenberg, one of the physicists who introduced the idea of colored quarks, showed that if CPT symmetry is broken, Lorentz symmetry must also be broken. Observations of antimatter are the best place to look for CPT violation — and if CPT violation is observed in experiments taking place at CERN and Fermilab we will know that Lorentz symmetry is also violated.

Kostelecký has developed a general framework for discussing these ideas. It is called the Standard Model Extension and it includes every possible way in which the fundamental particles of the Standard Model might interact so that *particle* Lorentz symmetry is violated while *observer* Lorentz symmetry is maintained. It includes theories, for example, in which neutrinos are the only particles that can undergo Lorentz-violating interactions, or theories in which photons can undergo Lorentz-violating interactions but the quarks and leptons cannot. The Standard Model Extension thus contains a large number of sectors in which Lorentz violation could occur. Most of these sectors have yet to be investigated with the required sensitivity — but several experiments are planned. It is possible — just possible — that these experiments will give us a glimpse of M-theory.

Chapter 11

THE EKPYROTIC SCENARIO: WHEN WORLDS COLLIDE

Q: What was God doing before He made heaven and earth?
A: He was preparing hell. . .for those who pry into mysteries.
Augustine, *Confessions, Book XI*

The concept of branes has been crucial in recent physics. We have seen how theorists learned that branes were an inevitable part of a consistent superstring theory and that they played a role that was just as important as strings. The study of branes and duality led to M-theory, an overarching 11-dimensional theory that has 11-dimensional supergravity and the 5 different superstring theories among its limits. The notion of branes was key in the explorations of black hole entropy, and it was a springboard for the large extra dimension theories of Arkani-Hamed, Dimopoulos, and Dvali. In the theory of Randall and Sundrum, we live on a brane. Thousands of papers have been published about the properties of branes, and the brane idea is one of the most fertile ideas around. To conclude our tour of the wild ideas of modern physics, let's look at one last idea that has been inspired by branes: the ekpyrotic universe.

In the standard model of cosmology, the Universe began 13.7 billion years ago in a singularity. It was in an exceptionally hot, exceptionally small, exceptionally dense state. Leaving aside the difficulties in understanding physics near a singularity — let alone what happened before the singularity — there was a problem with the early version of the Big Bang idea. If we look in different directions in space, we see the same things: the cosmic microwave background radiation is the same in all directions of the sky and the average distribution of matter, found in clusters of galaxies, is the same in all directions. On a large scale, the Universe appears to be smooth. The problem for the initial formulation of Big Bang theory is that, immediately after the Big Bang, widely separated parts of the expanding Universe would have been too distant from each other for light to have reached from one part to another. There would have been no way for these different parts of the Universe to equalize their temperature. When we look at the cosmic microwave background in patches of sky separated by more than a few degrees we *should* see different temperatures, because these patches were not in causal contact when radiation decoupled from matter, about 300,000 years after the Big Bang. These regions simply cannot "know" they should be at the same temperature. But we don't see a lumpy Universe. We see a smooth, equalized Universe. This is called the horizon problem, and it is one of the most difficult questions facing cosmologists.

The standard solution to the problem is called *inflation*, which as we saw on page 123 was first proposed by Alexei Starobinsky and Alan Guth. The theory was later developed by Andre Linde, then at the Lebedev Institute in Moscow, and Andreas Albrecht and Paul Steinhardt, who were both at the University of Pennsylvania. In the inflationary paradigm, the Universe underwent a brief period of exponential expansion just after the Big Bang. The entire visible Universe we observe now (and the entire Universe back in the microwave background era) was amplified from just a small, causally connected patch of the original Big Bang. That is why one side of the visible Universe has the same temperature and density as the other side. The idea also explains the tiny irregularities we observe in the background radiation: they were originally just tiny quantum fluctuations, which were made large by inflation. Inflation also explains other problems with the original Big Bang idea. There is, for example, the monopole problem: why do we not see the magnetic monopoles that a hot Big Bang would have produced in abundance? The answer is that inflation sweeps them all far away, beyond our horizon. Finally, inflation explains the so-called flatness problem: the mere fact that we are here, some 13.7 billion years after the Big Bang, means that the density of the Universe was equal to the critical density (or "flat") to within 1 part in 10^{15}. Had it been just a little more dense, the Universe would have recollapsed; had it been a little less dense, it would have blown apart. There seems to be no reason why the Universe should have been so close to the critical density immediately after the Big Bang — but a period of inflation would have made the Universe flat, regardless of the density with which it began.

For more than 20 years there has been no viable alternative to inflation as a theory that can solve the problems of the original Big Bang idea. There have been variants on the inflationary idea, to be sure, but they all maintain the central idea: we see a smooth Universe today because in the first fraction of a second of its life the Universe expanded exponentially. Inflation may well turn out to be the correct solution, and the model seems to work even better in a scenario with extra dimensions than it does in its original four-dimensions-only incarnation. But at least for now there is — perhaps — an alternative picture for the very early Universe. The picture was developed in 2001 by Burt Ovrut of the University of Pennsylvania, Neil Turok of Cambridge University, the Princeton theorist Paul Steinhardt (who was one of the developers of the inflationary model), and Steinhardt's graduate student Justin Khoury.

The alternative involves the world-as-a-brane scenarios. Suppose we live on a 3-brane floating in a bulk space with four spatial dimensions. And suppose there is another 3-brane, parallel to ours, also floating in the bulk. The branes could be boundary surfaces of the bulk space, like the two ends

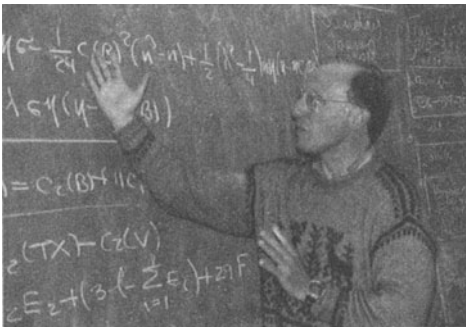


FIGURE 140 Burt Ovrut is a professor of theoretical high-energy physics at the University of Pennsylvania. His research is focused on M-theory and cosmology (and new mathematical methods in geometry that are required to study M-theory).

of a concertina. Suppose these boundary branes approached each other. What would happen? Well, branes are extremely massive so a moving brane contains huge amounts of kinetic energy. If branes collide, perhaps this kinetic energy can be liberated. So here then is an alternative explanation for the Big Bang. Initially our brane was empty and cold; the two boundary branes were moving toward each other in the unseen extra dimension of bulk space; the tremendous kinetic energy in the collision was converted into the familiar particles — the quarks, electrons, photons, and so on — that are confined to move along the three dimensions of the brane. Afterwards our brane expanded and continues to expand today. Everything that happened after the initial collision occurred just as it did in the standard Big Bang cosmology; the difference is that the Universe was never the small superdense object that is assumed in the standard Big Bang scenario.

FIGURE 141 Neil Turok

Khoury, Ovrut, Steinhardt, and Turok call this the *ekpyrotic universe* from the Greek word *ekpyrosis*, which means "conflagration." The ancient Stoics had a cosmological model in which the Universe began in a sudden burst of fire — rather like the more modern picture, which involves the collision of two branes.

There is an immediate advantage to the ekpyrotic scenario: although the energy involved is tremendous, the temperature during the collision is finite — there is no singularity. Furthermore, the successes of the inflationary model would seem to be reproduced. For example, the temperature of the collision is far too small to produce magnetic monopoles, which would explain why we do not see them; we do not need to postulate that an exponential inflation moved them from our sight. Furthermore, energy considerations ensure

that the collision would lead to us observing a flat Universe. And the Universe looks the same everywhere because the collision was between parallel branes; the collision occurred at the same time essentially everywhere on the brane.

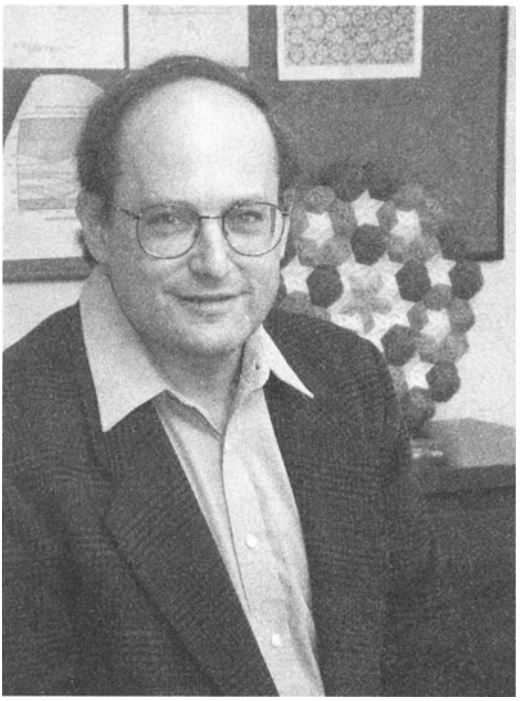


FIGURE 142 Paul Steinhardt is the Albert Einstein professor of science at Princeton University. He is one of the architects of the inflationary model in cosmology and has explored novel models of dark matter. Along with Khoury, Ovrut, and Turok, he introduced the ekpyrotic universe, an alternative to traditional Big Bang cosmology.

The ekpyrotic scenario also correctly explains the spectrum of the energy density fluctuations we observe in the cosmic microwave background. As the branes approach, quantum effects cause the brane to "ripple" along the extra dimension, so the collision occurs slightly earlier at some points than at others. Once the collision is complete, and the branes have separated, the rippling causes small temperature variations; these variations in turn seed galaxy formation and the fluctuations that we observe in the microwave background. The correct prediction of the spectrum of microwave background radiation fluctuations was considered to be a triumph for the inflation theory, but here is a model involving very different ideas that also seems to get the spectrum correct! Perhaps at last we have an alternative to inflation?

Well, there are problems with the scenario. For example, the ekpyrotic explanation of the observed smoothness of the Universe only works if the colliding branes are almost perfectly parallel; they have to clash like a pair of cymbals. If the branes collided at an angle to each other then

we would observe anisotropies in the sky. But why should the branes start out parallel? There seems to be no good answer to this question. There are problems, too, with explaining exactly how the tiny ripples in the fifth dimension are able to survive the collision and form the seeds of fluctuations in the microwave background. In any case, the ekpyrotic scenario may find it difficult to become the favored cosmological model. Inflation is built on well understood quantum field theory and has been studied for more than two decades. Ekpyrosis, however, is based on poorly understood ideas coming from M-theory and has undergone nothing like the scrutiny afforded inflation. Fortunately, we are blessed in living in an era in which cosmological questions are starting to be answered by observation: we may soon be able to decide whether ekpyrosis or inflation is correct.

The key to distinguishing between the two ideas is that inflation would magnify any primordial gravitional waves in such a way that the cosmic background radiation should now possess distinctive temperature fluctuations. In the ekpyrotic scenario there is no period of inflation, and so ekpyrosis predicts that there should be no such gravitational-wave effects in the background radiation. The *Planck* satellite is due to be launched in the spring of 2007, and it should have the capability of detecting a gravitational-wave signature in the microwave background. If it finds the signature then the ekpyrotic universe will be dead, and we will know that inflation is almost certainly the correct explanation for the horizon, flatness, and monopole problems mentioned above.

But until the observational data arrive, cosmologists will investigate the properties of an ekpyrotic universe in a little more detail. One of the variants of the initial idea that has been explored allows for a completely different view of our Universe.

In the traditional view, a view that is held by most cosmologists and is the accepted position of the Catholic Church, the Universe began at a particular instant. The Universe, in fact, began 13.7 billion years ago and asking what happened before then is a question that has no meaning. The ekpyrotic scenario allows for the return of an old idea: perhaps the Universe is infinitely old and had no beginning at all.

Steinhardt and Turok argued that dark energy — the mysterious energy of the vacuum that appears to be causing the expansion of the Universe to accelerate — could cause an attractive force in the bulk. The energy of the vacuum in the fifth dimension acts like a spring between the two boundary branes, so we can envisage the following picture. The two boundary branes are pulled towards each other by the vacuum energy, and they collide. The energy of the collision causes an event that observers that are bound to the brane (observers like us) call the Big Bang. The branes expand and cool, galaxies form on them, and eventually the energy of the vacuum on the

branes cause a gentle acceleration in the expansion. Meanwhile, however, the branes have bounced apart in the fifth dimension of bulk space. The attractive vacuum between the branes eventually halts their motion away from each other, and causes them to move towards each other. Eventually, the branes collide once again and a new Big Bang occurs. The cycle starts all over again. This collision takes place a *long* time after the earlier collision; by the time the branes collide again, they have been expanding for aeons. The branes are cool, and so the region of space that experiences a new Big Bang has the same temperature as the previous cycle. Each cycle is like the last (though specific details would be different from cycle to cycle). If we could observe the process from the bulk, we would see two branes constantly oscillating in the fifth dimension.

Perhaps, then, the Universe had no beginning. Perhaps it has always existed, along with its fellow Universe, and is condemned to oscillate forever. Occasionally, the two Universes collide in a fireball. And the result of one of those collisions is a world that is hospitable to life — and eventually to creatures that can ponder how the world was made.

The Anthropic Principle

Man is the measure of all things.
Protagoras

The quest that led physicists to M-theory was the quest for the unified theory of fundamental physics. Not just *a* unified theory, but *the* unified theory. The hope of most string theorists is that M-theory will have only *one* formulation that is mathematically consistent, and this unique formulation will contain within it the Standard Model of particle physics, general relativity, and those other features of the Universe that our experiments uncover.

Although it is too early to say for sure, this hope looks increasingly forlorn. As we saw in Chapter 7, there are many thousands of Calabi–Yau manifolds upon which a ten-dimensional superstring theory can be compactified. The geometry of each of these Calabi–Yau manifolds is described by hundreds of moduli, and every set of moduli gives rise to a different valid solution to the equations. Branes can wrap around the subspaces of a Calabi–Yau manifold in a bewildering variety of ways. All this means that there are perhaps more than 10^{100} different distinct vacua — different solutions — of string theory. Each of these solutions gives rise to different physics in the four-dimensional world we observe, *and each of these solutions is equally valid.* But, at least with our present level of understanding,

M-theory gives us no indication of which solution we should favor. We may be in the situation of possessing a unique unified theory — but one that has over 10^{100} equally valid solutions, with no way of choosing the solution that describes the Universe we actually observe.

There is a way around this problem: to invoke the anthropic principle. This approach is unpalatable to many physicists, but increasingly it seems we may be forced to accept some form of this approach. The anthropic principle stems from the observation that many aspects of our Universe seem to be "fine-tuned" so that life can occur and flourish. If the cosmological constant were just a little larger, or if the weak force were just a little stronger, or if neutrinos had just a little more mass — well, we would not be here to observe the outcome. A Universe in which fundamental parameters were just a little different would be a lifeless Universe. Well, suppose that the Universe is much, much bigger than the part we observe. Suppose that this stupendously large Universe contains all possible vacuum structures of M-theory; each vacuum structure would correspond to a different "pocket" Universe. There would be more than 10^{100} such "pocket" Universes, each with its own laws of physics, its own version of the Standard Model, even its own number of spatial dimensions. We simply happen to live in one of those Universes with properties that allow our existence.

Coming Attractions

World is crazier and more of it than we think.
Louis MacNeice, *Snow*

By the time this book is published, exotic physics may have been discovered in tabletop gravity experiments, in cosmic ray observatories, in gravity wave detectors, or in analyses of existing collider data. These experiments and analyses may give us our first glimpse of physics that is "out of this world." But perhaps they will find nothing that cannot be explained by the existing twin pillars of physics: relativity and quantum field theory. What then? Well, then we will have to wait until spring 2007. That is the expected start-up date for the Large Hadron Collider.

The LHC will collide beams of protons with a total energy of 14 TeV. Each beam will consist of a train of proton bunches — about 3000 bunches, each of which will be about 7.5 centimeters in length. After each train has passed, there will be a pause lasting 3 microseconds, during which time no beam is present. More than 1300 extremely powerful superconducting magnets will steer the proton beams around the 27-kilometer-long tunnel that once housed the LEP collider. And at carefully chosen points in the

tunnel, the proton beams will be smashed together in the most energetic collisions ever engineered.

Many pieces of evidence suggest that at energies around the TeV scale, new physics should come into play. And the LHC will collide sufficient numbers of particles at sufficiently high energies that physicists will be able to investigate the TeV scale in detail.

So what might they find?

The first sighting of interest should surely be the Higgs boson. We saw in an earlier chapter how the usual incarnation of the Standard Model requires the existence of the Higgs boson. Theoretical considerations, and data from existing collider experiments at places like LEP, imply that the Higgs mass must be in a range that it easily accessible to the LHC. If necessary, the LHC will simply step through the possible range of values that the Higgs mass can take, and it will find the elusive boson. That should be the first discovery.

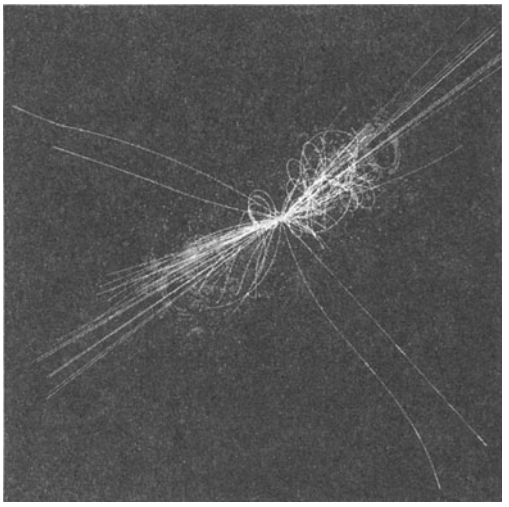


FIGURE 143 This is another computer simulation of what a detector might see if a Higgs boson decays into four muons. Two of the muons emerge from the center and move towards the upper-left part diagram; two of the muons emerge from the center and move towards the lower-right part diagram. The Higgs might be the first new particle to be discovered at the LHC.

Another discovery should be clear evidence for supersymmetry. If the supersymmetric partners of the gluons and quarks — the gluino and squark — have a mass no greater than about 2 TeV, then the LHC will find them. Although 2 TeV is much heavier than any known particle, the LHC will have the energy to generate them. And for years physicists have studied the sorts of signal that the production of these supersymmetric particles would create. If the LHC produces supersymmetric particles, we know what to be looking for.

Supersymmetry has been an important part of the thinking of particle physicists for many years. However, we saw earlier in this chapter how one

of the main problems that we can solve by invoking supersymmetry — the gauge hierarchy problem — can be solved in other ways. The existence of large extra dimensions that gravity alone can feel would explain the relative weakness of gravity and solve the hierarchy problem. If this scenario is correct, the LHC could become a black hole factory. It is a tantalizing prospect.

We saw how the hierarchy problem can be solved even more naturally in a scenario with a warped fifth dimension. If models based on this idea are correct, then the LHC should produce Kaluza–Klein excitations of the graviton. These new particles would be free to move into the fifth dimension, taking energy with them. We would be sending particles into a hidden extra dimension!

There are a whole host of possibilities that theorists have dreamed up involving extra dimensions, only some of which we have described, and soon the LHC will put these ideas to the test.

Suppose, though, that the LHC finds none of these things. What then? Well, the non-discovery of the Higgs would be a bigger surprise than its discovery. With sufficient ingenuity, though, theorists would be able to generate extensions to the Standard Model that did not require the Higgs. (Non-Higgs theories, indeed, have been studied for many years.) The non-discovery of supersymmetry, or failure to find evidence for larger dimensions, would be more disappointing. We might then be asked to accept on faith that M-theory describes our Universe, but that the extra dimensions of space are curled up at the inaccessible Planck scale of 10^{-35} meters and that nature somehow breaks supersymmetry at an energy that is inaccessible to our colliders. That would be unfortunate: we would remain — for who knows how long — in the position of having a theory that we did not know how to test. We might be forced to accept an anthropic explanation: out of the vast landscape of valid M-theory solutions, we happen to inhabit the solution that allows for our existence.

But there is another possibility. Perhaps the LHC will not find the Higgs or supersymmetry or extra dimensions. But perhaps it will find something completely unexpected! The LHC will be entering unexplored territory, and perhaps it will find dragons. This is, after all, why we build machines like the LHC: to investigate the unknown, not just to confirm what we already think we know. That is the most exciting prospect of all — an experimental discovery that will send theorists back to their chalkboards and force them to dream up even wilder ideas to explain the Universe that our instruments reveal to us.

And we would once again be searching for answers that are truly out of this world.

GLOSSARY

Abelian group: A *group* of transformations in which the end result of a product of transformations does not depend upon the order in which they occur. In a non-Abelian group, on the other hand, the order in which transformations take place is important.

AdS/CFT correspondence: A *symmetry* that relates a theory of *gravity* in an *anti-de Sitter spacetime* (AdS) to a *conformal field theory* (CFT) without gravity on the boundary of that *spacetime*. There is thus a deep connection between gravity and field theory, the exact nature of which is still being researched.

annihilation: A process in which a particle and its *antiparticle* collide and destroy each other, thereby creating new particles. For example, an electron and an antielectron can mutually annihilate to produce two photons.

anomaly: An anomaly occurs in *relativistic quantum field theory* when a conservation law arising from a *symmetry* in the underlying theory is destroyed by quantum corrections. The requirement that a theory must be free from anomalies imposes tight restrictions on its structure.

anti-de Sitter spacetime: A highly symmetrical solution to the equations of *general relativity*, which describes an empty Universe that possesses a negative *cosmological constant*.

antiparticle: A particle and its antiparticle possess exactly the same *rest mass*, but otherwise have "mirror-reversed" properties. For example, a particle has an opposite electric charge to its antiparticle. Some particles, such as the *photon*, serve as their own antiparticle.

asymptotic freedom: In *quantum chromodynamics*, the weakening of the *color* interaction between *quarks* and *gluons* as the distance between them decreases.

baryon: A subatomic particle that is a bound state of three *quarks*. Baryons have half-integer *spin*. *Protons* and *neutrons* are the most common type of baryon, though over one hundred different type baryons have been discovered in collider experiments.

Big Bang: Our expanding Universe is thought to have begun 13.7 billion years ago in an event called the Big Bang; immediately after the Big Bang, the Universe is thought to have been incredibly hot and dense.

black hole: An object with a gravitational field that is so intense that nothing — not even light — can escape once the hole's *event horizon* has been crossed.

boson: A particle with integer *spin*. Fundamental bosons include the *photon*, *gluon*, *W boson*, and *Z boson*, all of which are spin-1 particles; the *graviton* is a spin-2 boson. These fundamental bosons can thus be thought of as "particles of force." Another type of boson, the spin-0 *Higgs boson*, is believed to be the mechanism by which fundamental particles acquire mass; the Higgs (as of April 2004) has not yet been observed in experiments.

BPS state: A particular type of quantum state possessing *supersymmetry*, which has properties that can be uncovered by *symmetry* arguments.

brane: A type of object in *M-theory*. A *p*-brane has *p* spatial dimensions; so a 0-brane is a particle, a 1-brane is a string, a 2-brane is a membrane, and so on. In some theories, our Universe may be a 3-brane embedded in a higher-dimensional space.

branon: An extremely speculative quantum particle. If our Universe is a *brane* embedded in higher dimensions, the brane can fluctuate in those dimensions. The rules of quantum mechanics tell us that such fluctuations would manifest themselves as particles called branons.

bulk: In some theories, multidimensional *branes* are embedded in a higher-dimensional space called the bulk. Gauge fields might be confined to the brane, whereas *gravity* would be free to propagate throughout the bulk.

Calabi–Yau space: A particular type of space into which the six extra dimensions of *superstring theory* may undergo *compactification*.

CERN: Conseil Européen pour la Recherche Nucléaire — the European Organization for Nuclear Research — is the world's largest particle physics laboratory. It is located near Geneva, on the border between France and Switzerland. The *Large Electron–Positron Collider* at CERN closed in 2000, making way for the powerful *Large Hadron Collider*, which will become operational in 2007.

charge: A quantity carried by a particle that determines its participation in an interaction process. For example, if a particle possesses electric charge then it participates in the *electromagnetic interaction*; if a particle possesses *color* charge then it participates in the *strong interaction*.

chirality: The property of "handedness." The *weak interaction* is chiral: it distinguishes between a left-handed and a right-handed *fermion*.

classical physics: Physics prior to the introduction of the quantum principle. In classical physics, which incorporates Newtonian mechanics, quantities such energy are viewed as a continuum. In *quantum mechanics*, such quantities come in discrete amounts. Classical physics works extremely well in the everyday world; when physicists probe small distance scales, however, quantum mechanics must be used.

closed string: In *string theory*, a loop of string that has no free ends. The *graviton* is interpreted to be a specific vibrational pattern of a closed string.

cold dark matter: Observations have revealed that halos of invisible matter surround galaxies. This dark matter adds up to many times the mass of visible matter. The temperature of dark matter is a measure of how fast the particles in it are moving. Cold dark matter refers to particles moving much slower than the speed of light. (Hot dark matter refers to particles that move at speeds close to that of light.)

color: A type of *charge* carried by *quarks* and *gluons*. There are three different types of color charge, usually labelled as red, green, and blue. An isolated color charge is never seen in nature; we only ever observe color-neutral combinations of quarks. So a *baryon* is a combination of three quarks (one red, one green, one blue), and a *meson* is a combination of a quark and an antiquark (with a color and its anticolor).

compactification: The extra spatial dimensions demanded by *string theory* are conventionally thought to be curled up into unobservably small spaces (or compactified).

confinement: A free isolated *quark* (or *gluon*) has never been observed. Instead, colored quarks and gluons are confined within colorless *hadrons*. There is as yet no exact proof that the theory of colored quark and gluon interactions — *quantum chromodynamics* — exhibits confinement; recently, however, physicists have proved that supersymmetric quantum chromodynamics possesses confinement.

conformal field theory (CFT): A type of *relativistic quantum field theory* that looks the same at all energy scales. In *string theory*, the vibrations of strings can be described by a two-dimensional conformal field theory that lives on the surface of the string's *world sheet*.

conformal symmetry: Conformal transformations are those that leave the angles between directions invariant.

conservation law: When an interaction between particles takes place, the value of certain quantities (for example, electric charge, energy, and momentum) are the same before and after the interaction. The conservation laws are some of the most profound statements we can make about our Universe, and they ultimately derive from the symmetries of the Universe.

cosmic ray: A high-energy particle — often a *proton*, or an atomic nucleus — that hits Earth's atmosphere from outer space.

cosmological constant: A term that can be added to the original equations of *general relativity*. Nowadays we interpret the term as the energy density of the vacuum, and calculating its value is one of the great challenges of *M-theory*. These theories imply that the cosmological constant is large; observation, though, tells us it is small.

coupling constant: A parameter that describes the strength of a given interaction.

de Sitter spacetime: A highly symmetrical solution to the equations of *general relativity*, in which the Universe contains no matter but has a positive *cosmological constant*. Such a Universe expands at an exponentially increasing rate.

dilaton: A hypothetical *scalar field*, which occurs naturally in *Kaluza–Klein theory* and other theories employing higher spatial dimensions.

duality: A *symmetry* in which two seemingly different theories produce identical physical consequences.

ekpyrotic scenario: A speculative theory in which our Universe began through the collision of two parallel *branes*.

electromagnetic interaction: The interaction that occurs between objects due to their electric *charge*. The interaction is mediated by the exchange of *photons*. The *relativistic quantum field theory* of the interaction is called *quantum electrodynamics*.

electron: A type of fundamental particle (a *lepton*). Electrons carry one unit of negative electric *charge* and possess a *rest mass* of 0.511 *MeV*. They make up the outer shells of atoms.

electron Volt (eV): The energy gained by an *electron* when it is accelerated through an electrical potential of 1 Volt. The binding energy on an electron in an atom is of the order of 1 eV; this relatively low energy makes it easy to produce free electrons.

electroweak interaction: At high energies, the *electromagnetic interaction* and the *weak interaction* can be described by a single electroweak theory. The theory of the electroweak interaction is a key part of the *Standard Model*.

entropy: A measure of the amount of disorder in a physical system. When systems interact, the total entropy cannot decrease.

equivalence principle: Over a small enough region of observation, it is impossible to distinguish between the effects of a gravitational field and those of accelerated motion. The equivalence principle means that all observers, no matter what their state of motion, can argue that they are at rest as long as they take an appropriate gravitational field into account. The equivalence principle was the foundation upon which Einstein constructed his theory of *general relativity*.

event horizon: The area around a *black hole* at which the escape velocity is equal to the speed of light. The horizon acts as a one-way membrane in space: once an object has passed the event horizon, it can never escape from the hole's gravitational pull.

exclusion principle: Identical *fermions* are forbidden from sharing the same quantum state. (*Bosons*, by contrast, do not obey an exclusion principle; any number of bosons can be in the same quantum state.)

extended supergravity: A type of *supergravity* theory that possesses more *supersymmetry* transformations than the single ($N = 1$) transformation of the original supergravity theory. There are eight supergravity theories, ranging from $N = 1$ (the original theory, which has one *gravitino*) to $N = 8$ (which has eight gravitinos). The particle content of extended supergravity theories is precisely constrained by *symmetry*.

extremal black hole: A *black hole* possessing the minimum mass consistent with its charge. Such a black hole is closely related to a *BPS state*.

Fermilab: The American National Accelerator Laboratory in Batavia, IL. It is named after the famous physicist Enrico Fermi. At present, accelerators at Fermilab produce the most energetic particle collisions in the world; this distinction will pass to the *Large Hadron Collider*, when the *CERN* machine becomes operational.

fermion: A particle with a half-integer *spin*. *Leptons* and *quarks* both have spin-$\frac{1}{2}$ and so are fermions. Fermions can be thought of as "particles of matter."

Feynman diagram: A graphical representation of particle interactions. Feynman diagrams define a recipe for calculating the likelihood of a given quantum process taking place.

flavor: A characteristic that distinguishes different types of *quark*. There are six quark flavors: up, down, strange, charm, top, and bottom.

gauge theory: A *relativistic quantum field theory* based upon gauge *symmetry*. Two important gauge symmetries are the *Abelian group* $U(1)_{EM}$ and the non-Abelian group $SU(3)_c$. The former describes the interaction of a *photon* with an electrically charged particle, the latter describes the interaction of a *gluon* with a color charged particle (both quarks and gluons carry the color charge). These are unbroken gauge symmetries. The group $SU(2)_L \times U(1)_Y$, which is spontaneously broken, describes the *electroweak interaction*. A gauge symmetry implies the existence of massless gauge bosons; the spontaneous breaking of a gauge symmetry gives mass to some of the gauge bosons.

general relativity: Einstein's extension of *special relativity*, which shows that *gravity* can be thought of as the curvature of *spacetime*.

giga electron Volt (GeV): One billion (10^9) *electron Volt*. Using Einstein's *mass–energy equivalence* ($E = mc^2$), the mass of the *proton* is just under 1 GeV.

global symmetry: A *symmetry* of a system that holds if the same transformation is made at all *spacetime* points.

glueball: A colorless bound state of two or more *gluons*.

gluino: The supersymmetric partner of the *gluon*.

gluon: A spin-1 *boson* that is the mediator of the *strong interaction*. Since gluons carry the *color* charge, they too feel the strong force and thus interact with themselves.

grand unified theory (GUT): A theory that unites the *strong interaction* and the *electromagnetic interaction*. The merging of the forces takes place only at very high energies, which so far have been inaccessible to experiment. A generic prediction of grand unified theories is that the *proton* should decay, and physicists have been searching — so far without success — for signs of this.

graviton: The massless, electrically neutral, spin-2 quantum particle of the gravitational field. In *string theory*, the graviton is a *closed string*.

gravitino: The supersymmetric spin-$3/2$ partner of the *graviton*.

gravity: A force that acts between objects according to their mass. In Einstein's picture, gravity arises from the distortion of *spacetime* caused by an object's mass–energy. In a quantum picture, the gravitational interaction between objects would arise through the exchange of spin-2 *gravitons*.

group: If a collection of elements possesses four simple properties when an operation acts on any pair of them, the elements are said to form a group under that operation. The four properties are: closure (i.e. if a and b are elements of the group, then the operation on the pair a and b produces an element that is also in the group), associativity (i.e. the order in which two successive operations are applied to the elements is not important), identity (i.e. there is an element e of the group that has a neutral action, so that the operation on the pair a and e always leaves a unchanged), and inverse (i.e. for every element there is another element such that the operation acting on this pair of elements produces the identity). The importance for physics is that *symmetry* transformations form groups and can thus be studied using *group theory*.

group theory: A branch of mathematics dealing with *symmetry*.

hadron: A composite particle made up of *quarks*. There are two types of hadron: a *baryon* is a three-quark state (qqq) and a *meson* is a quark–antiquark state ($q\bar{q}$).

helicity: The projection of a particle's spin along its direction of motion. A particle's helicity is either left-handed or right-handed, depending upon whether its spin is in the direction of motion or against it.

heterotic string: A *closed string*, which supports right-moving oscillations that resemble those of a type II superstring and left-moving oscillations that are the same as a bosonic string. There are two distinct types of heterotic string theory: one based on the gauge group SO(32) and one based on the product E(8) × E(8). These are two of the five consistent *superstring theories*.

hierarchy problem: The problem of understanding why the *electroweak* energy scale (roughly 1 *TeV*) is so far removed from the *Planck scale* (roughly 10^{16} TeV). Quantum effects should drag the electroweak scale up to the Planck scale. Attempts to solve the hierarchy problem have produced some interesting ideas in theoretical physics.

Higgs boson: A particle associated with a spin-0 field that is conjectured to pervade the Universe. It plays an important role in theories involving spontaneous symmetry breaking, such as the theory of the *electroweak interaction*. Fundamental particles are believed to generate their mass through interactions with the Higgs field.

Higgsino: The supersymmetric partner of the *Higgs boson*.

high-energy physics: Another term for particle physics. To investigate the constituents first of atoms, then nuclei, then protons, physicists were driven to higher and higher energies (which enabled them to investigate smaller and smaller distances).

holographic principle: The idea that a quantum theory of *gravity* must be able to describe all phenomena happening within a volume by degrees of freedom that exist on the surface of that volume.

Kaluza–Klein theory: A type of quantum mechanical theory in which extra spatial dimensions are supposed to exist. The non-observation of the extra dimensions is explained by *compactification*: we assume the dimensions are curled up to small sizes.

Large Electron–Positron Collider (LEP): A particle accelerator, 26.7 kilometers in circumference and 100 meters underground, situated at *CERN*. It ran for 11 years. Until its closure in 2000 to make way for the *Large Hadron Collider*, experiments at LEP confirmed the predictions of the *Standard Model* to a high degree of precision.

Large Hadron Collider (LHC): A particle accelerator at *CERN* that will bring protons and ions into head-on collisions at higher energies than ever achieved before. The first operation of the collider is planned for spring 2007.

length contraction: According to *special relativity*, a moving object will appear to be foreshortened along its direction of motion.

lepton: A type of fundamental particle that does not possess *color* charge and thus does not interact via the *strong interaction*; it interacts only through the *electromagnetic interaction* and the *weak interaction*. There are three families of lepton: the *electron* and its *neutrino*, the *muon* and its neutrino, and the *tau* and its neutrino.

Lie group: A continuous *symmetry* of certain types of mathematical equation. The properties of all possible Lie groups have been well studied by mathematicians. Lie groups play an important role in theoretical studies of particle physics.

lightest supersymmetric particle (LSP): Most models of *supersymmetry* possess a quantity that suggests the least massive supersymmetric particle must be stable: no lighter particles exist into which it can decay. A candidate for the LSP is the *neutralino*.

local symmetry: A *symmetry* of a system that holds if a different transformation is made at different *spacetime* points. A symmetry can hold under such restrictive conditions, so long as a compensating gauge field is introduced.

Lorentz invariance: A *symmetry* that says the laws of physics manifest themselves identically in all inertial frames of reference.

magnetic monopole: A hypothetical magnetic *charge*, not yet observed in experiments. The magnetic field produced by a monopole is similar to that produced by the end of a very long, infinitely thin coil of wire (known as a solenoid). There are general reasons for supposing that magnetic monopoles must exist in a wide class of *grand unified theories*.

mass–energy equivalence: According to special *relativity*, a measured quantity of energy is equivalent to a measured quantity of mass (given by the equation $E = mc^2$, where E is energy, m the equivalent mass, and c the speed of light).

mega electron Volt (MeV): One million (10^6) *electron Volt*. Using Einstein's *mass–energy equivalence* ($E = mc^2$), the *rest mass* of the *proton* is about 938 MeV and the mass of the *electron* is about 0.511 MeV.

meson: A subatomic particle that is a bound state of a *quark* and an antiquark pair. The pions are the lightest mesons, and they consist of up and down quarks and antiquarks. Other important mesons include the K mesons and the B mesons.

moduli: Parameters that determine the size and shape of the various dimensions of a compactified *Calabi–Yau space*. The moduli are *scalar* fields; potentially, such scalar fields could play an important role in cosmology and *high-energy physics*.

momentum: The product of an object's mass and its velocity.

M-theory: A framework that unites the five consistent 10-dimensional *superstring theories* with 11-dimensional *supergravity*. Theorists have uncovered many isolated bits of information about M-theory, but a complete understanding of the theory still seems far from their grasp.

muon: A type of *lepton*, more massive than an *electron* but less massive than a *tau*.

neutralino: The lightest electrically neutral supersymmetric particle. It would be a spin-$1/2$ particle — a quantum combination of the *photino*, *Zino*, and *Higgsino*. The neutralino is a candidate for *cold dark matter*. As with all supersymmetric partners to the usual particles, the neutralino has not yet been observed.

neutrino: An electrically neutral *lepton*. Neutrinos were once thought to be massless; it now seems likely that they possess mass. (Although, if they do, they are extremely light.) There is a neutrino for each of the charged leptons: an *electron* neutrino, a *muon* neutrino, and a *tau* neutrino.

neutron: An electrically neutral *baryon* consisting of a ddu triplet of *quarks*. The neutron is a basic constituent of the atomic nucleus.

Noether's theorem: Informally stated, Noether's theorem says that to every continuous *symmetry* of a theory there corresponds a *conservation law* and vice versa.

open string: In *string theory*, a string with two free ends. The ends can be glued to an object called a *brane* and are responsible for many of the properties of a brane. (Open strings are not responsible for gravity, however; gravity arises through the vibration of *closed strings*.)

parity: A quantum number that characterizes the *symmetry* of the *wavefunction* of a particle (or a system of particles) under a discrete transformation. For example, C parity characterizes the symmetry of a wavefunction when the particle is replaced by its antiparticle; P parity characterizes the symmetry of a wavefunction when the system is reflected in a mirror.

particle accelerator: A machine, such as those at *CERN* or *Fermilab*, that accelerates beams of particles to close to light speed and then collides the beams. By studying the debris of the collisions, physicists can probe the fundamental structure of matter.

perturbation theory: A method for solving a difficult problem by finding successively more accurate approximations. For example, the true numerical answer to a problem might be 10.576. The zeroth-order approximation in perturbation theory might be 10; the first-order perturbation might be 0.5; the second-order perturbation might be 0.08; and thus we converge upon the answer.

photino: The supersymmetric partner of the *photon*.

photon: A spin-1 *boson* that is the mediator of the *electromagnetic interaction*. A photon does not carry electric *charge*, so does not itself feel the electromagnetic interaction.

Planck area: An area that is simply the square of the *Planck length*. Its value is $1.64 \times 10^{-69}\,\mathrm{m}^2$. According to the *holographic principle*, nature can encode only about 1 bit of information for every Planck area.

Planck constant: The quantum of action — the fundamental quantity that sets the scale for quantum mechanical effects. It gives the ratio of the energy of a *photon* to its frequency; numerically, it is equal to $6.62620 \times 10^{-34}\,\mathrm{J\,s}$. It is symbolized by h.

Planck length: The "natural" unit of length, in that it is defined in terms of the fundamental constants of nature. It is equal to $\sqrt{Gh/c^3}$, where h is the Planck constant, G is Newton's gravitational constant, and c is the speed of light. It has the value 4.05×10^{-35} meters. At the Planck length, quantum effects dominate.

Planck mass: The "natural" unit of mass, in that it is defined in terms of the fundamental constants of nature. It is equal to $\sqrt{hc/G}$, where h is the Planck constant, G is Newton's gravitational constant, and c is the speed of light.

Planck scale: If we extrapolate the history of the Universe backwards in time, cosmology tells us the Universe becomes denser and hotter and the relevant distance scales become shorter. If the distance scales become short enough (of atomic dimensions or smaller), we must use *quantum mechanics* to explain phenomena. As we extrapolate backwards to the Big Bang, we eventually reach a time where a full quantum theory of *gravity* is needed. This is the Planck era, and the corresponding scales of distance, energy, and time constitute the Planck scale.

positron: The antielectron. The positron thus has one unit of positive electrical *charge*, but exactly the same *rest mass* as the *electron*.

proton: A positively charged *baryon* consisting of a uud triplet of *quarks*. The proton is a basic constituent of the atomic nucleus.

quantum chromodynamics (QCD): The *relativistic quantum field theory* describing the interactions of *quarks* via the exchange of *gluons*.

quantum electrodynamics (QED): The *relativistic quantum field theory* describing the interactions of electrically charged particles via the exchange of *photons*.

quantum mechanics: The physics that applies on very small scales. The main feature that distinguishes it from classical mechanics is that various quantities — such as energy, electric charge, momentum, and angular momentum — come in discrete amounts called quanta.

quark: A type of fundamental particle that possess *color* and thus interacts via the *strong interaction* as well as the *electromagnetic interaction* and the *weak interaction*. Quarks come in six types (or *flavors*): up, down, strange, charm, top, and bottom. The (u, d) quarks form one family, the (s, c) quarks form another family, and the (t, b) quarks form a third family.

radion: The quantum particle of the radion field, which is a *scalar field* that is conjectured to exist in *bulk* space and acts to stabilize the position of *branes* in the two-brane *Randall–Sundrum scenario*. The radion is a massive particle; some studies suggest that, if the particle exists, it may be observable at the *Large Hadron Collider*.

Randall–Sundrum scenario: The suggestion that our Universe is a 3-brane embedded in a 5-dimensional *spacetime*, with the extra spatial dimension of the *bulk* being warped. If there is another 3-brane, separated from ours by a small distance in the extra dimension, and if gravity is localized to this brane, then the warp factor would ensure that gravity is weak on our brane. Thus this scenario addresses the *hierarchy problem*. A later version of the scenario postulates the existence of only one brane in a warped 5-dimensional spacetime.

relativistic quantum field theory: The marriage of *special relativity* and quantum field theory, in which quantum fields are the ultimate reality. Fundamental particles are the quanta of the fields and can be created or destroyed. The theory is thus used to describe the creation and destruction of particles in elementary particle interactions.

relativity: The notion that space and time are relative rather than absolute concepts. Relativistic ideas date back to Galileo, but the term usually refers to Einstein's two theories of special and general relativity. The theory of special relativity is a fundamental theory of the spacetime properties of all physical processes. It is based on two principles. First, no experiment conducted in a closed physical system can determine whether the system is at rest or moving with constant speed in a straight line. Second, there is a limiting speed of propagation of signals. The theory of general relativity is based on the *equivalence principle*; it leads to the notion that the gravitational force is communicated via the curvature of *spacetime*.

renormalization: A process in *relativistic quantum field theory* whereby quantities that are infinite — such as the mass and charge of an electron — can be adjusted so that the infinities do not show up in the final computations. Some field theories, like *quantum electrodynamics* and *quantum chromodynamics* are renormalizable and thus yield finite answers to questions. The quantum field theory of *gravity* is non-renormalizable, and yields infinite answers.

rest mass: The intrinsic mass of a fundamental particle; it is the mass we measure if the particle is at rest relative to us. If a particle has zero rest mass, then it moves at the speed of light. Any particle with a non-zero rest mass cannot move at the speed of light.

scalar field: A spin-0 field. The *Higgs boson* is the quantum of a scalar field; so is the hypothetical *radion*.

S-duality: A *symmetry* whereby a *string theory* with a large *coupling constant* and a string theory with a small coupling constant give rise to identical physics.

selectron: The supersymmetric partner of the *electron*.

singularity: A point in *spacetime* at which the laws of physics break down; it can be thought of as a puncture in the fabric of spacetime.

SLAC: The Stanford Linear Accelerator Center. Using high-energy electron beams at the SLAC *particle accelerator*, physicists found the first evidence for the *quark* structure inside protons, and discovered the charm quark and the *tau* lepton.

slepton: The supersymmetric partner of a *lepton*. For example, the *selectron* is the supersymmetric partner of the *electron*; the *sneutrino* is the supersymmetric partner of the *neutrino*.

sneutrino: The supersymmetric partner of the *neutrino*.

soliton: A localized disturbance in a continuous medium that can propagate over long distances with no change in its shape or amplitude. In many ways, a soliton behaves like a particle.

spacetime: According to the theory of *special relativity*, time and space are not independent of each other; instead, we live in a four-dimensional spacetime — a combination of space (three dimensions) and time (the fourth dimension).

special unitary group (of order N), SU(N): A *Lie group*. It describes an abstract *symmetry* that can rotate N objects into each other. The groups SU(2) and SU(3) play an important role in particle physics.

spin: A quantum mechanical property of particles that is similar in some ways to the familiar concept of spin. Quantum particles possess an intrinsic spin, and in units of the *Planck constant* it can be an integer spin (in which case the particle is a *boson*) or a half-integer spin (in which case the particle is a *fermion*).

spontaneous symmetry breaking: This occurs when a system is described by equations that possess a certain *symmetry*, but the stable solutions to those equations lack the symmetry. The symmetric states are unstable, and so there is a spontaneous transition to a stable — but asymmetric — state. For example, a needle that balances on its tip has rotational symmetry — but this is an unstable situation. The needle will inevitably fall, thus spontaneously breaking the rotational symmetry.

squark: The supersymmetric partner of a *quark*.

Standard Model: A set of field theories, based on the gauge group $SU(3) \times SU(2) \times U(1)$, which describes the three fundamental non-gravitational interactions (the *electromagnetic interaction*, the *weak interaction*, and the *strong interaction*).

string theory: Also called *superstring theory*. By postulating that the fundamental objects of nature are one-dimensional strings, rather than zero-dimensional point particles, it is possible to develop a consistent theory that incorporates a quantum theory of *gravity* and the three fundamental non-gravitational interactions in a unified framework.

strong interaction: The interaction that occurs between objects due to their *color* charge. The interaction is mediated by the exchange of *gluons*. The strong force binds quarks together to form *hadrons*; an indirect effect of this force is that it binds *protons* and *neutrons* together to form atomic nuclei. The *relativistic quantum field theory* of the interaction is called *quantum chromodynamics*.

supergravity: When *supersymmetry* is made a *local symmetry*, general *relativity* — or *gravity* — appears automatically. The original supergravity theory contains only two fields — the *graviton* and the *gravitino* — but *extended supergravity* theories provide more realistic descriptions of nature. Supergravity theories are low-energy limits of *string theories*.

Super-Kamiokande: A large water-filled detector that is used to study *neutrinos*, *cosmic rays*, and certain phenomena — such as *proton* decay and the existence of *magnetic monopoles* — predicted by *grand unified theories*.

superstring theory: See *string theory*.

supersymmetry (SUSY): A conjectured *symmetry* that relates *fermions* (particles having half-integer *spin*) with *bosons* (particles having integer spin).

symmetry: Invariance under transformation. A system thus possesses a symmetry if it is unchanged when it is transformed in some way.

tachyon: A hypothetical particle that moves at faster-than-light speeds. If a theory contains a tachyon, the theory is likely to be inconsistent.

tau: The most massive of the three families of *lepton*.

T-duality: A *symmetry* whereby a *string theory* compactified on a small circle and a string theory compactified on a large circle give rise to identical physics.

tera electron Volt (TeV): One trillion (10^{12}) electron Volts; equivalently, it is 1000 *GeV*. This is a convenient unit of energy for the world's most powerful *particle accelerators*, which can produce beams of *protons* with an energy of about 1 TeV. The *Large Hadron Collider* will collide protons at a total energy of 14 TeV.

thermodynamics: The study of heat flow in a physical system; it is concerned with quantities like work, energy, and *entropy*.

time dilation: In special *relativity*, the slowing of time for observers in relative motion.

uncertainty principle: A principle of *quantum mechanics* that states there are pairs of quantities (such as position/momentum and energy/time) that can never be measured with complete precision. We can measure one of the pair of quantities with ever-increasing precision, but that is at the expense of ever-increasing ignorance of the other quantity. The uncertainty principle gives rise to quantum effects that are quite impossible in *classical physics*.

vector field: A spin-1 field. The electromagnetic field is a vector field.

virtual particle: An "internal" particle in a *Feynman diagram* that does not possess the mass it would possess if it were a real, free particle. A virtual *photon*, for example, can possess mass; this is in contrast to a real photon, which is massless. The emission and absorption of virtual particles, in a way described by *relativistic quantum field theory*, forms the basis of all fundamental particle interactions.

wavefunction: A function of position and time that is a solution to certain types of wave equation. The square of the wavefunction that is the solution to the Schrödinger wave equation — the fundamental equation in *quantum mechanics* — has a special meaning: it gives a probability density.

wavelength: The distance between successive crests, troughs, or identical parts of a wave.

wave–particle duality: A fundamental feature of *quantum mechanics*, whereby an object like an *electron* exhibits both wavelike and particle-like properties.

W boson: A fundamental spin-1 particle, the exchange of which (along with the *Z boson*), mediates the *weak interaction* of *leptons* and *quarks*. The W boson can possess an electric charge of +1 or −1; it has a *rest mass* of 81.8 *GeV*.

weak interaction: The interaction that occurs between objects due to their weak *charge*, which is possessed by all fundamental particles; the weak interaction is the only one that is known to affect the *neutrino*. The interaction is mediated by the exchange of the *W boson* and the *Z boson*. At high energies the interaction is described by *electroweak theory*, which treats the *weak interaction* and the *electromagnetic interaction* in a unified way.

weak hypercharge: The *electroweak interaction* is based upon a spontaneously broken SU(2) × U(1) gauge symmetry. Here, U(1) is the group of weak hypercharge, which is simply related to electric *charge*.

weak isospin: The *electroweak interaction* is based upon a spontaneously broken SU(2) × U(1) gauge symmetry. Here, SU(2) is the group of weak isospin, which can be thought of as a type of *charge*.

weak mixing angle: A fundamental parameter of the theory of the *electroweak interaction*. In the *Standard Model*, the value of this parameter must be put in by hand; *grand unified theories* can predict the value of this parameter.

WIMP: Weakly interacting massive particle. These conjectured subatomic particles, which are a candidate for explaining the "dark matter problem" (the problem that most of the mass of the Universe is in some as-yet undetected form), would have little interaction with normal matter except for gravitational attraction.

Wino: The supersymmetric partner of the *W boson*.

world line: The one-dimensional line that a point particle traces out as it moves through *spacetime*.

world sheet: The two-dimensional surface that a string sweeps out as it moves through *spacetime*.

X boson: A hypothetical gauge *boson* of SU(5) *grand unified theory*. The X boson can change a *quark* into a *lepton*, and vice versa. The X bosons carry *color* and have an electric *charge* of $-4/3$.

Y boson: A hypothetical gauge *boson* of SU(5) *grand unified theory*. The Y boson can change a *quark* into a *lepton*, and vice versa. The Y bosons carry *color* and have an electric *charge* of $-1/3$.

Z boson: A fundamental spin-1 particle, the exchange of which (along with the *W boson*), mediates the *weak interaction* of *leptons* and *quarks*. The Z boson is electrically neutral; it has a *rest mass* of 92.6 *GeV*.

Zino: The supersymmetric partner of the *Z boson*.

BIBLIOGRAPHY

Since this is intended for a general audience, I have not listed the original references to the research mentioned in this book. Such papers are usually intensely mathematical. However, the reader who is interested in reading the original papers should visit http://arxiv.org, the e-print service for physics and mathematics; searches can be made by author and date.

The world's major particle physics laboratories are keen to explain the research they are engaged in, and their websites contain much useful and accessible news and information. Visit, for example, http://public.web.cern.ch/public/ (the CERN website) and http://www.fnal.gov/ (the Fermilab website). A particularly useful website for particle physics is http://www.interactions.org/cms/.

The volumes listed below are recommended for the reader who wants to learn more about some of the concepts discussed in this book; all of them are accessible to the non-mathematical reader.

Abbott, E. A. 1962. *Flatland: A Romance of Many Dimensions.* Oxford: Blackwell.
The original popularization of higher dimensions. This is an old reprinted edition; newer editions are also available.

Adair, R. K. 1987. *The Great Design.* Oxford: OUP.
A good introduction to the ideas of relativity and particle physics.

Barrow, J. D. and Tipler, F. J. 1986. *The Anthropic Cosmological Principle.* Oxford: OUP.
The main source of reference if you are interested in anthropic arguments in physics.

Bloom, A. 1991. *The Republic of Plato.* 2nd ed. New York: Basic Books.
Probably the best English translation of Plato's famous work.

Bryson, B. 2003. *A Short History of Nearly Everything.* New York: Broadway Books.
This book contains a large amount of science that is well-organized and delivered in Bryson's usual excellent style.

Close, F. E. 2000. *Lucifer's Legacy.* Oxford: OUP.
An accessible book about symmetry — and asymmetry — in physics.

Close, F. E., Marten, M., and Sutton, C. 1987. *The Particle Explosion.* Oxford: OUP.
A beautifully illustrated introduction to the particle physics of the Standard Model.

Epstein, L. C. 1985. *Relativity Visualized.* San Francisco: Insight Press.
An entertaining introduction to relativity.

Froggatt, C. and Nielsen, H. 1991. *Origin of Symmetries.* Singapore: World Scientific.
Contains many of the key papers on supersymmetry and superstrings to which I have referred in this book. It is very heavy going, of course, but it also contains lucid descriptions of the importance of symmetry in physics.

Geroch, R. 1978. *Relativity from A to B.* Chicago: University of Chicago Press.
A thorough, accessible introduction to the concepts of general relativity.

Greene, R. 1999. *The Elegant Universe.* New York: Vintage.
Easily the best introduction to the ideas of superstrings and one of the finest popular physics books in recent years.

Kaku, M. 1994. *Hyperspace.* Oxford: OUP.
A clear, accessible account of higher dimensions and superstring theory.

Lederman, L. M. and Schramm, D. N. 1989. *From Quarks to the Cosmos.* New York: Scientific American Library.
A beautifully illustrated account of particle physics.

Lilley, S. 1981. *Discovering Relativity for Yourself.* Cambridge: CUP.
One of the best introductions to the principles of relativity.

Smolin, L. 2000. *Three Roads to Quantum Gravity.* Oxford: OUP.
Includes a lucid description of loop quantum gravity.

Weinberg, S. 1992. *Dreams of a Final Theory.* New York: Pantheon Books.
A book about the search for the fundamental laws of nature, by one of the great theoretical physicists.

Wilczek, F. and Devine, B. 1987. *Longing for the Harmonies.* New York: Norton.
Includes a beautiful description of QCD by one of the creators of the theory.

Will, C. M. 1993. *Was Einstein Right?* Oxford: OUP.
Describes the many experimental tests of general relativity.

INDEX

Abrikosov, A. A., 191
Adelberger, E. G., 265, 277
AdS/CFT correspondence, 236, 240–242
Akulov, V. P., 128
Albrecht, A. J., 279
allegory of the cave, 4, 5
Anderson, C. D., 69
Anderson, P. W., 103
anthropic principle, 284
anti-de Sitter spacetime, 237, 240, 272
Antoniadis, I., 245, 249, 255
Arkani-Hamed, N., 249–257, 259, 262, 263, 265, 266, 268–270, 274, 278
Ashtekar, A. V., 155
Auger, P. V., 261
axion, 108

Bach, J. S., 14
Bachas, C. P., 255
Banks, T., 205
Bars, I., 211
baryon, 65
Bekenstein, J. D., 218–221, 223, 224, 230
Bergshoeff, E. A., 209
Berners-Lee, T., 58
beta decay, 68
Big Bang, 1, 46, 57, 107, 246, 259, 279, 280
black hole, 207, 215–217, 219–223
black hole entropy, 223
Bogomolny, E. B., 193
Boltzmann, L., 217, 218, 223, 226–228
Bose, S. N., 62
boson, 62–64, 72
Bousso, R., 235
BPS state, 193, 214, 223
Brandenberger, R. H., 200
brane, 202, 204–212, 243, 246, 248, 250, 255–257, 259, 262, 267, 269, 271–274, 278–283
brane scan, 210
branon, 269
Brink, L., 165
Brookhaven National Laboratory, 82
Brout, R., 103
bulk, 248–250, 255–257, 259, 262, 266, 267, 271, 273, 279, 280, 282, 283

Calabi, E., 171
Calabi–Yau manifold, 171, 173, 176, 199, 224, 249, 283
Callan, C. G., 224
Candelas, P., 172
Carnot, S. N. L., 217
Cavendish, H., 263
CERN, 52, 53, 56, 58, 106, 277
charge, 59–61, 76, 80, 81, 86, 88–91, 99, 115, 124, 136, 215

chirality, 20, 98
chirality problem, 152
Clausius, R. J. E., 217, 223, 226
Cleland, A. N., 41
color, 89–92, 94, 95, 110, 190, 277
compactification, 145–152, 166, 167, 170–179, 196–200, 204, 205, 211, 222–224, 238, 248, 249, 269–271, 283
confinement, 67, 94, 188–191, 193, 194, 242
conformal symmetry, 161
conservation law, 19, 20
cosmic ray, 259
cosmological constant, 239
coupling constant, 81, 87, 88, 92–96, 101, 110, 115, 117, 118, 125, 128, 130, 136, 176, 177, 190, 192, 195, 196, 201, 202, 204
Cowan Jr, C. L., 68
CP symmetry, 108
CPT symmetry, 277
Cremmer, E., 151, 202

D-brane, 213, 214, 221–223, 235, 243, 247
Dai, J., 212
dark energy, 282
dark matter, 131, 255, 256
Das, S. R., 224
de Sitter spacetime, 237
de Sitter, W., 237
Descartes, R., 21
Dienes, K. R., 255
dilaton, 146, 148, 149, 176, 177, 195, 266
Dimopoulos, S., 249–257, 259, 261–263, 265, 266, 268–270, 274, 278
Dirac field, 60
Dirac, P. A. M., 45, 69, 184, 185, 209
Dirichlet boundary condition, 212
Dirichlet, J., 213
duality, 182, 184, 188, 190, 192–195
Dudas, E., 255
Duff, M. J., 202, 211
Dvali, G., 249–257, 259, 262, 263, 265, 266, 268–270, 274, 278

Einstein, A., 2, 3, 8, 21–23, 30, 31, 34, 35, 37, 38, 44, 47, 62, 132, 137, 141–144, 182, 202, 215, 216, 237, 239, 274
ekpyrotic universe, 278, 280–282
electromagnetic field, 44
electromagnetism, *see* QED
electron, 52, 61, 65, 68, 69, 81, 86–88, 129, 161
electron neutrino, *see* neutrino
electron volt (eV), 52
electroweak theory, 136
Englert, F., 103
entropy, 217–219, 226
equivalence principle, 35–38, 182
Escher, M. C., 14, 238
event horizon, 216, 218–220, 223, 230–232, 236
exceptional groups, 28, 111
exclusion principle, 62, 63, 90, 137
extremal black hole, 222

Fedorov, E. S., 13
Feng, J. L., 261
Fermi, E., 62, 68, 111, 112
Fermilab, 52, 56, 72, 277
fermion, 62–64
Ferrara, S., 135, 138
Feynman diagram, 85
Feynman, R. P., 3, 44, 85, 86, 88
Fischler, W., 205
flavor, 65, 66, 70, 72, 73, 75, 95, 96
Freedman, D. Z., 135, 138
Fritzsch, H., 123

Gabor, D., 233
Galilei, G., 21–24
Galois, E., 27
gauge boson, 100–103, 110, 114, 116, 124, 133–136, 154, 164, 178, 185, 248

gauge field, 83, 90, 134, 135, 172, 185, 249
gauge invariance, 60, 85, 89, 99, 106, 110, 133, 136, 149, 150, 153, 167, 168, 178
gauge theory, 84, 85, 89, 94, 99, 100, 109, 111, 114, 138, 146, 150, 166, 167, 171, 176, 178, 179, 193, 196, 214, 225, 236, 240–244, 250, 255
gedanken experiments, 2, 4
Gell-Mann, M., 65, 70, 90
general relativity, *see* relativity
Georgi, H., 114, 118, 123
Gherghetta, T., 255
Giddings, S. B., 261, 268
Glashow, S. L., 99–101, 106, 114, 169, 174, 175
Gliozzi, F., 164
Global Positioning System, 7, 30–33, 42, 200
glueball, 76, 244
gluino, 129, 133
gluon, 90–92, 95, 110, 129, 133
Goldberger, W. D., 273
Golfand, Y. A., 127
Gorkov, L. P., 191
Goto, T., 156, 160
grand unified theory (GUT), 114–121, 123–126, 153, 185, 188, 255
gravitational redshift, 37
gravitino, 135
graviton, 47, 135, 154, 163, 257, 266, 267
gravity, 3, 7, 28, 34–38, 46, 47, 79, 112, 118, 119, 126, 130, 135–137, 141–143, 146–155, 158, 163–172, 177, 178, 202, 206, 213–217, 220, 221, 226, 231, 232, 235–241, 243, 244, 248–258, 262–266, 269–275, 284, 286
Green, M. B., 165, 167, 169, 181, 211, 225
Greenberg, O. W., 277
Greene, B. R., 199
Gross, D. J., 93, 169
group theory, 25, 27
Gubser, S. S., 241
Gursey, F., 124
Guth, A. H., 123, 279

hadron, 65
Harvey, J. A., 169
Hawking radiation, 224
Hawking, S. W., 218–221, 223, 224, 230, 262
Heckel, B. R., 265, 277
Heraclitus, 6
heterotic string, *see* superstrings
hierarchy problem, 125, 126, 128, 133, 138, 254, 255, 257, 259, 269, 270, 272, 274, 286
Higgs boson, 102–104, 124–126, 134, 274, 285
Higgs, P. W., 102
Higgsino, 131
Hořava, P., 201, 248, 249, 272
holographic principle, 232, 235, 240, 241, 243
Horowitz, G. T., 172, 224
Howe, P. S., 211
Hughes, J., 209
Hull, C. M., 195
Huxley, T. H., 3, 121

Inami, T., 211
inflation, 123, 279, 281, 282
isospin, 100

Julia, B., 151, 202

Kaluza, T. F. E., 142–145, 149, 152
Kaluza–Klein theory, 145, 149, 152, 162, 171, 176, 196, 209
Kennedy–Thorndike experiment, 276
Khoury, J., 279, 280
Kikkawa, K., 196
Klebanov, I. R., 241

Klein, O., 143–149, 152, 245
Knobel, R., 41
Kostelecký, V. A., 275–277

Landsberg, G. L., 261, 268
Large Electron–Positron Collider (LEP), 106, 266, 269, 285
Large Hadron Collider (LHC), 52, 58, 107, 133, 134, 176, 261, 268, 272, 274, 284, 286
Laser Interferometer Gravitational Wave Observatory (LIGO), 263, 264
Lederman, L. M., 69, 72
Leigh, R. G., 212
leptons, 65, 80
Lie groups, 27, 124
Lie, M. S., 27, 28
lightest supersymmetric particle (LSP), 131, 256
Likhtman, E. P., 127
Linde, A., 279
Liu, J. T., 209
loop quantum gravity, 154, 155
Lorentz symmetry, 23, 24, 59, 274–277
Lorentz symmetry violation, 275–277
Lorentz, H. A., 23

M-theory, 181, 182, 250, 272, 275
magnetic monopole, 182–185, 188
Maldacena, J. M., 224, 235, 236, 238, 240–243, 272
Mandlestam, S., 188
Martinec, E. J., 169
Mathur, S. D., 224
matrix model, 206
Maxwell, J. C., 39, 44, 101, 142, 182, 184
Meissner effect, 191
Meissner, W. H., 191
meson, 65, 161
Michelson–Morley experiment, 276
Miller, D. J., 104, 105
Minkowski, H., 22, 143
Minkowski, P., 123
mirror manifold, 199
mirror symmetry, 199
moduli, 177
monopole, 191
Montonen, C., 190, 192
muon, 69, 72
muon neutrino, *see* neutrino

Nakano, T., 70
Nambu, Y., 156, 160
Neddermeyer, S. H., 69
neutralino, 131, 132
neutrino, 68, 69, 81, 257
  electron neutrino, 68, 69
  muon neutrino, 69
neutron, 66, 69, 91
Neveu, A., 128, 161
Newton, I., 8, 38, 253, 257, 263, 266
Nielsen, H. B., 156, 160
Nishijima, K., 70
Noether's theorem, 17, 19, 23, 60
Noether, E., 19, 161
Nordström, G., 143, 149

Ochsenfeld, R., 191
Olive, D. I., 164, 190, 192
Ovrut, B. A., 279, 280

parity, 20, 108, 277
particle accelerator, 52
Pauli, W., 20, 68, 132, 148
Peccei, R. D., 108
pentaquark, 67
Perl, M., 71, 72
perturbation theory, 87, 88, 92–94, 136, 158, 160, 167, 176, 189, 192, 193, 196, 202, 204, 242
photino, 129, 131
photon, 62, 129
Pierre Auger Cosmic Ray Observatory, 260–262
pion, 65, 66, 70, 161
Planck constant, 40
Planck length, 47, 148, 247

Planck mass, 168, 217
Planck satellite, 282
Planck scale, 249
Plato, 4–6, 10
Platonic solids, 10
Plesser, M. R., 199
Poincaré symmetry, 127
Poincaré, J. H., 127
Polchinski, J. G., 209, 212, 247
Politzer, H. D., 93
Polyakov, A. M., 185–188, 241
Pope, C. N., 211
positron, 45, 75, 76
Prasad, M. K., 193
Price, J. C., 265, 266
proton, 50, 52, 53, 65, 66, 69, 70, 91, 93, 161
proton decay, 117

quantum chromodynamics (QCD), 81, 89–93, 95, 109, 110, 136, 154, 163, 188, 190, 194, 241–244
quantum electrodynamics (QED), 80, 84–94, 104, 109, 136, 137, 154
quantum field theory, 44
quantum mechanics, 43
quarks, 65–67, 69, 80, 185
Quinn, H. R., 108, 118

radion, 273
Ramond, P. M., 124, 128, 161
Randall, L. J., 270–274, 278
Randall–Sundrum scenario, 269–274
Reines, F., 68, 71
relativistic quantum field theory, 43, 44
relativity
  general theory of, 3, 7, 8, 32, 34, 123, 132, 142, 146, 177, 182, 215, 216, 237, 239
  special theory of, 3, 8, 17, 21, 23, 32, 34, 43
renormalization, 88, 89, 93, 94, 136
rest mass, 59
Rohm, R., 169
Rosenbaum, T. F., 41
Rovelli, C., 155
R-parity, 131
Rubakov, V. A., 245

S-duality, 195, 196, 201
Sakharov, A., 120
Salam, A., 101, 102, 106, 128, 169
Samuel, S., 275
Saprophilous, M. E., 245
Scherk, J., 151, 163, 164, 202
Schrödinger, E., 42
Schwartz, M., 69
Schwarz, J. H., 128, 161, 163–165, 167, 169, 181, 211, 225
Schwarzschild, K., 215
Schwinger, J. S., 88, 99
Scott Russell, J., 185, 186
Seiberg, N., 192–194, 235
selectron, 129
Sen, A., 192
Sezgin, E., 209, 211
Shannon, C. E., 226–228
Shapere, A. L., 261
Shenker, S. H., 205
Sikivie, P., 124
singularity, 216
slepton, 134
Smolin, L., 155
sneutrino, 129
SO(10), 123
soliton, 185, 186, 188
Sommerfield, C. M., 193
Sorkin, R. D., 230
special relativity, *see* relativity
spin, 61
squark, 133, 134
Standard Model, 7, 8, 79, 80, 83, 85–96, 101–115, 125, 128, 129, 138, 146, 150, 151, 153, 154, 158, 160, 166, 169, 173, 178, 179, 188, 194, 206, 248, 255, 256, 266, 269, 276, 285, 286

Standard Model Extension, 277
Starobinsky, A. A., 123, 279
Steinberger, J., 69
Steinhardt, P. J., 279, 280, 282
Stelle, K. S., 211
Strathdee, J., 128
string tension, 157
string theory, *see* superstrings
Strominger, A. E., 172, 211, 221–224
strong interaction, 70, 81, 83, 85, 89–95
SU(5) grand unification, 114, 115, 121, 123
Sundrum, R., 270–274, 278
superconductor, 191
supergravity, 135, 137, 153
supernova, 258
superstrings, 153–180
  heterotic, 169, 170, 196, 198, 201, 204, 205, 211, 213, 248
  type I, 196, 201
  type IIA, 201, 202, 204
  type IIB, 195, 196, 201, 238
supersymmetry (SUSY), 127–129, 131–134, 153, 193, 204, 285
Susskind, L., 156, 160, 205, 226, 235, 243
symmetry, 10–12, 14

T-duality, 196
tabletop gravity experiments, 263
tachyon, 164
tau, 72
tau neutrino, *see* neutrino
Tevatron, 52, 72, 133, 176, 266–268
thermodynamics, 217, 219, 228–232
Thomas, S., 261, 268
Thomson, J. J., 65
't Hooft, G., 93, 104, 185–188, 190, 194, 225, 226, 232, 235, 242, 243
time reversal, 277
Tomonaga, S-I., 88
Townsend, P. K., 195, 209
Turok, N. G., 279, 280, 282

uncertainty principle, 40, 42, 47, 119, 144, 201, 220

Vafa, C., 200, 221–223
van Nieuwenhuizen, P., 135, 138
Veneziano, G., 156
Volkov, D. V., 128
von Neumann, J., 227

W boson, 95, 96, 100, 101, 103, 119
wave–particle duality, 41
wavefunction, 42, 45, 63
weak hypercharge, 100
weak interaction, 70, 73, 81, 95, 96, 99
weak isospin, 99, 100
weak mixing angle, 115
Weeks, J. A., 10
Weinberg, S., 3, 44, 101, 102, 106, 118, 169
Wess, J., 128
Wheeler, J. A., 38
Wigner, E. P., 59, 62
Wilczek, F., 93
Wino, 129
Wise, M. B., 273
Witten, E., 150, 152, 165, 172, 181, 193–196, 201–204, 235, 241, 248, 249, 272
world line, 22, 157, 161
world sheet, 157, 160, 162
World Wide Web, 58

X boson, 114, 119

Y boson, 114
Yamasaki, M., 196
Yau, S.-T., 171
Yoneya, T., 163

Z boson, 95, 96, 103
Zino, 131
Zumino, B., 128
Zweig, G., 65